

EMMA: an AMR cosmological simulation code with radiative transfer

Dominique Aubert^{1*}, Nicolas Deparis¹ and Pierre Ocvirk¹

¹*Observatoire Astronomique de Strasbourg, CNRS UMR 7550, Université de Strasbourg, Strasbourg, France*

1 September 2015

ABSTRACT

EMMA is a cosmological simulation code aimed at investigating the reionization epoch. It handles simultaneously collisionless and gas dynamics, as well as radiative transfer physics using a moment-based description with the M1 approximation. Field quantities are stored and computed on an adaptive 3D mesh and the spatial resolution can be dynamically modified based on physically-motivated criteria.

Physical processes can be coupled at all spatial and temporal scales. We also introduce a new and optional approximation to handle radiation : the light is transported at the resolution of the non-refined grid and only once the dynamics have been fully updated, whereas thermo-chemical processes are still tracked on the refined elements. Such an approximation reduces the overheads induced by the treatment of radiation physics. A suite of standard tests are presented and passed by EMMA, providing a validation for its future use in studies of the reionization epoch.

The code is parallel and is able to use graphics processing units (GPUs) to accelerate hydrodynamics and radiative transfer calculations. Depending on the optimizations and the compilers used to generate the CPU reference, global GPU acceleration factors between x3.9 and x16.9 can be obtained. Vectorization and transfer operations currently prevent better GPU performances and we expect that future optimizations and hardware evolution will lead to greater accelerations.

Key words: cosmology: dark ages, reionization, first stars - methods: numerical

1 INTRODUCTION

Starting in the 70s, numerical simulations have been part of the astrophysicists tool kit to investigate regimes where non-linearities, strong coupling between different physics and multi-scale processes are dominant. The study of structure formation in a cosmological context is an archetypal example of such intricacies and has thus driven the rise and the growth of the so-called ‘cosmological simulations’ for decades now. These simulations have successively included collisionless dynamics and hydrodynamics, and implements routinely sub-resolution physics such as star formation or feedback from supernovae to model the galaxy formation process.

Recently, the challenges encountered in galaxy formation theory and the near advent of new observatories capable of investigating the Universe at redshifts $z > 6$ (such as JWST or SKA) produced an interest toward the study of the reionization epoch. Interesting in its own sake, as a great cosmic transition produced by the first sources and capable of reionizing the Universe, this epoch will be studied *in-situ*

and will provide insights on the initial stages of the structure formation process (see e.g. reviews by Barkana & Loeb 2001; Pritchard & Loeb 2012). Reionization is also thought to provide keys to understand the current state of galaxies, with the rise of a UV background capable of suppressing star formation in light objects (see e.g. Gnedin 2000; Hoeft et al. 2006; Finlator et al. 2011; Wise et al. 2012). Consequently, numerical simulations of the reionization started to appear 15 years ago to assess this process and prepare the advent of observational constraints (see Trac & Gnedin 2011 for a review).

The common feature of such simulations is the inclusion of radiative transfer physics, to model the impact of the UV photons emitted by the first sources. Quite demanding in terms of computing resources, radiative transfer is often included as a post-processing step on outputs of simulations to reduce this cost : absorbents are static and ad-hoc recipes are included to model the negative feedback of radiation on sources (see e.g. Ciardi et al. 2003; Iliev et al. 2006; McQuinn et al. 2007; Mellema et al. 2006; Zahn et al. 2007; Baek et al. 2010; Chardin et al. 2012; Ocvirk et al. 2013; Paardekooper et al. 2013; Ocvirk et al. 2014; Zawada et al. 2014). Nevertheless, radiative hydrodynamics codes started to be imple-

* E-mail: dominique.aubert@astro.unistra.fr

mented (like recently e.g. Finlator et al. 2011; Rosdahl et al. 2013; Pawlik et al. 2015). They cope with the reduced time scales (and therefore increased CPU cost) produced by radiation physics using massive parallelism (see e.g. Trac & Cen 2007; Norman et al. 2015), methodological solutions (such as implicit solvers like e.g. González et al. 2007; Norman et al. 2015) or specific physical regimes (such as a reduced speed of light like e.g. Gnedin 2014; Rosdahl et al. 2013). Whatever will be the solution, radiation will surely become an additional standard component of future cosmological simulation codes. This evolution is greatly illustrated by the number of participating codes to comparison projects such as Iliev et al. (2006) and Iliev et al. (2009).

In this context, a new cosmological simulation code is presented here, **EMMA**¹, that includes collisionless physics, hydrodynamics and radiative transfer. It builds upon the experience gathered with previous codes such as **ATON** (Aubert & Teyssier 2008, 2010) or a particle-mesh N-Body only code (Aubert et al. 2009). Its ambition is to be able to tackle structure formation experiments with a special emphasis on the influence of radiation during the reionization epoch. **EMMA** is an adaptive mesh refinement code, parallel and with GPU-driven acceleration on a selection of its sub-modules, mainly the physics engines. As such, it shares a number of features with **ATON** and extend its field of application to coupled radiative hydrodynamics and to high resolution through mesh refinement.

The full methodology of **EMMA** is first described: details on how data is managed, how the physics is solved and how parallelisation is implemented are presented. Then we present a selection of validations tests, from pure dark matter experiments to more complex situations with coupled physics. Finally we discuss the performances and the potential future developments of **EMMA**.

2 METHODOLOGY

2.1 Adaptive Mesh implementation

EMMA relies on a grid-based description of the physical quantities and implements adaptive mesh refinement (AMR) techniques to increase dynamically its spatial resolution. The latter is typically of the order of a few cells, for all the different physics presented here, whether it is set by the smoothing of the gravitational force or the smearing of hydrodynamical shocks or ionization fronts. AMR permits to increase this resolution with a moderate cost in terms of memory consumption by providing finer meshes only at selected locations.

Several AMR implementation exists and **EMMA** adopts a Fully Threaded Tree (FTT, Khokhlov (1998)) description of the data, sharing this core feature with other codes such as **ART** (Kravtsov et al. 1997) or **RAMSES** (Teyssier 2002). In this framework, a fundamental grid divides a 3D space (usually cubic but not necessarily) in $2^{3\ell_c}$ cells where ℓ_c designates its refinement level and where the cells are arranged in a Cartesian manner. For instance $\ell_c = 7$ corresponds to a fundamental 128^3 grid where each direction is sampled along

128 points. Hereafter this fundamental coarsest grid will be referred as the *base* grid and ℓ_c is the *base* level.

These base level cells are the roots of oct-trees that fully describe the refined geometry of space. If a cell is refined, it points towards an octal structure (or *oct*) that in turns points toward eight additional cells belonging to the $\ell_c + 1$ level. This hierarchy can be recursively maintained until a satisfying resolution (i.e. refinement level) is achieved. Conversely, base level cells can also be grouped in octs, that belong to coarser cells of level $\ell_c - 1$. Recursively, this process can be repeated until a single $\ell = 0$ cell is obtained. This specific cell is therefore the root of the global AMR structure that samples the 3D space where the numerical experiments take place.

In practice, the data is organized using octs as the fundamental structures where a level ℓ oct O^ℓ contains the following informations:

- the level ℓ ,
- 8 cells c_i^ℓ with $i \in [0, 7]$,
- a pointer toward its parent cell $c^{\ell-1}$,
- 6 pointers toward the 6 Cartesian neighbors cells of $c^{\ell-1}$,
- 2 pointers toward the next and the previous octs in memory that belong to the same level ℓ . Note that the arrangement of octs in memory is not related to spatial distribution of the octs in the computational domain. With this feature, octs are stored as a *doubly linked list*.

A cell c_i^ℓ , contains the following informations:

- its index $i \in [0, 7]$,
- the physical data at this location (density, potential, pressure, ionized fraction, etc...),
- a pointer toward an oct $O^{\ell+1}$ if it is refined,

Octs are mainly used for data exploration and management. They are the mandatory intermediates in order to probe the neighbors of a given cell. They are also the intermediate structure to scan all the cells at a given ℓ (i.e. at a given spatial resolution) as they store the pointers toward the previous and next member of the set of all the O^ℓ . This scanning operations are performed through the doubly linked list.

Cells are mainly used to store physical data. They also store a pointer toward an oct if they are refined but that is all the relational information that they possess. Any query on a neighbor cell must be passed through their parent oct. This description reduces greatly (by a factor 8 typically) the number of relational pointers that would be required if each cell possessed its own set of neighbors: such an organization takes advantage of the fact that all the cells of an oct share a significant fraction of neighbors, at the cost of some operations to retrieve them via the octs.

Since geometrical relations between neighbors are explicit, the mesh can be arbitrarily refined without any constraint on the geometrical strategy of this refinement. Also there is no simple mapping between the position of an oct in memory and its geometrical location. As a consequence, the fully threaded tree can be extremely versatile and follow closely any physical feature that requires high resolution at the cost of storing additional information.

In refined grids, resolution jumps require special care. In particular these interface are source of errors and inac-

¹ Electromagnetisme et Mecanique sur Maille Adaptative

curacies that must be controlled as much as possible. One standard requirement that must be enforced in the current description of AMR data is the fact that resolution jumps cannot be greater than unity for two neighbor cells. Hence the six neighbor pointers of an oct necessarily exist. However it also puts constraints on the way cells are refined and the procedure takes place as follow, in that order:

(i) an oct containing a cell marked for refinement or already refined must be maintained to ensure that resolution jumps are smaller than two. Its parent cell is therefore marked for refinement.

(ii) if a cell is marked for refinement, so must be its 26 neighbors. It ensures a smooth transition between low and high resolution regions by creating a buffer of high resolution cells at the interface and also prevents large resolution jumps.

(iii) if a cell satisfies some user-defined criterion, it is marked for refinement. This criterion can be physically motivated (based on physical values or gradients) or based on location (for zoom simulation for instance).

In practice, step (i) is applied to all the cells that belong to a given level in a single pass. Afterwards the step (ii) criterion is likewise applied to all the cells of that level, then step (iii), resulting in 3 successive passes on all the cells of a given level. The marking procedure can be repeated an arbitrarily number of times, especially to increase the thickness of the buffer regions at the transition between regions at 2 different resolutions. In practice, this procedure is applied twice, similarly to **RAMSES** or **ART**. It ensures that a cell that is refined on a physical basis (criterion 3 above) will be produced with an outer layer of 2 neighbor cells at the same resolution. Once this marking has been completed, cells can be refined by creating a new oct to be attached to it or coarsened by suppressing its child oct. When refined, all the relations must be computed while the physical quantities are straightforwardly injected from coarse values.

2.2 Solvers

EMMA tracks the evolution of 3 ‘fluids’, coupled to each other. First a collisionless fluid, sampled by particles in a standard particle-in-cell description. It is aimed at modeling the dynamics of stars and non-baryonic dark matter. Second, the code solves the Euler equations to describe the gas dynamics. We chose to follow **RAMSES** by using a piecewise linear method ‘a la’ MUSCL-Hancock driven by HLLC Riemann solvers. Third, **EMMA** models the propagation of radiation using a moment description of the radiative transfer (RT hereafter) equation with the M1 closure relation. Preliminary atomic processes are also included to describe the cooling and ionization that takes place on top of the adiabatic evolution of the gas. Finally, **EMMA** deals with cosmological settings through the use of supercomoving coordinates, which are briefly discussed.

2.2.1 Collisionless dynamics

Collisionless dynamics are handled through a Monte-Carlo sampling of phase space using particles (see e.g. Hockney & Eastwood (1981) for a reference). Each particle consists in a structure of data that contains:

- (i) its fundamental properties: mass, 3D position $\mathbf{r} = x, y, z$ and velocity $\mathbf{v} = v_x, v_y, v_z$,
- (ii) the level of the cell it belongs to,
- (iii) 2 pointer towards the next and the previous particle that belong to the same cell.

Additionally, if collisionless dynamics is included, a non-refined cell contains a pointer toward the first particle it contains. If a cell is split, the particles of the coarse cell are passed to the newly created cells. If an oct is destroyed, all the particles are assigned to the parent cell. By means of linked lists, the code can therefore scan all the particles of a given cell.

Each particle is evolved in phase space thanks to the usual Newton’s equations :

$$\frac{d\mathbf{v}}{dt} = -\nabla\phi(\mathbf{r}), \quad (1)$$

$$\Delta\phi(\mathbf{r}) = 4\pi G\rho(\mathbf{r}), \quad (2)$$

where $\rho(\mathbf{r})$ stands for the 3D total matter density and $\phi(\mathbf{r})$ is the associated gravitational potential. Updating the phase space coordinates of a particle therefore implies an evaluation of the density and the potential fields on the refined mesh.

The density is computed in all cells of the AMR grid using a standard cloud-in-cell (CIC) interpolation scheme. In practice it is performed level by level, according to the time stepping procedure described in section 2.3. If an unsplit cell c_i^ℓ contains particles, they are assigned to cells of the same level ℓ according to the CIC scheme. Furthermore if a neighbor cell is split, the contribution of the particles in its $\ell + 1$ cells is also taken in account to compute the density of c_i^ℓ . Conversely, if a neighbor cell is coarser and unsplit, its particles will also contribute to the density of c_i^ℓ . In these two situations, the CIC extent of the particles will correspond to the level ℓ resolution, resulting in a CIC density equals to the one that would be obtained from a uniform grid at level ℓ . Finally if c_i^ℓ is split, its density is computed by averaging the 8 values of its child cells.

The density being available, the potential is obtained by solving the Poisson equation. In the case of **EMMA**, the solution is obtained by means of relaxation techniques. They allow for a great flexibility (notably compared to FFTs based approaches) in dealing with complex domain geometries, arbitrary boundary conditions and grids with multiple resolutions. One can use the simple Jacobi iteration formula where the estimation $p + 1$ for the potential in the cell of Cartesian coordinates (i, j, k) can be computed from a previous estimation p via:

$$\phi_{i,j,k}^{p+1} = \frac{\phi_{i+1}^p + \phi_{i-1}^p + \phi_{j+1}^p + \phi_{j-1}^p + \phi_{k+1}^p + \phi_{k-1}^p}{6} \quad (3)$$

$$+ \frac{4\pi G\Delta x^2 \rho}{6} \quad (4)$$

The convergence rate of this process is slow, especially to achieve convergence on scales comparable to the total volume. As can be seen in eq. 4, information can only be propagated from one cell to the next one, making such formula inefficient at determining the low frequency modes of the solution. It can be marginally accelerated by means of over-relaxation or by using the Gauss-Seidel iteration technique, analog to Jacobi but where new estimates of the potential

ϕ^{p+1} are used on the RHS of equation 4 as soon as they are available instead of ϕ^p .

Greater acceleration rates can be obtained by using a multigrid algorithm. The first stage is a ‘smoothing operation’ where a simple relaxation formula such as eq. 4 is used to remove spurious high frequency features in the current estimation of the potential, i.e. in the regime of scales it is the most efficient. Then this estimate of ϕ^p is coarsened through a restriction operation and applied as a first guess for a new relaxation stage, which is more efficient at low resolution for two reasons : first the number of sampling points is reduced (typically by a factor of 8 in 3D) and second, the convergence of low frequency modes is increased as their effective scale is reduced relative to the influence radius of eq. 4. Once this low resolution solution is converged, it is interpolated back (also called prolongation step) and after a few relaxation steps at full resolution a converged solution is obtained.

This two level-process (between ℓ and $\ell - 1$) can be recursively applied to $\ell - 1$ and $\ell - 2$ and so on. In **EMMA**, we typically proceed until $\ell = 2$ (i.e. using a 4^3 field) to achieve a 10^{-3} convergence of the solution with only ~ 5 iterations at full resolution. We take advantage of the AMR structure to store residuals and perform the calculations on coarse levels. Prolongation is performed via linear interpolation, whereas restriction is done by averaging on the 8 cells of an oct.

For fine levels with $\ell > \ell_c$, the potential is computed by simple relaxation, using a red-black Gauss-Seidel smoother, with boundary conditions provided by the $\ell - 1$ cells that surround high resolution patches. However, the initial potential evaluation is obtained by interpolation of the coarse solution. Therefore, convergence is quickly obtained from this first value that is already close to the correct one. For **EMMA**, convergence at the 10^{-3} level are obtained after 10 iterations in a typical cosmological simulation. Fine cells at resolution jumps must compute the potential using high resolution values that do not exist on the coarse side of the jumps. In this case, the potential is interpolated at the requested location from its coarse value.

The potential being available, the acceleration field $\mathbf{f} = -\nabla\phi$ is computed by simple derivation and is interpolated back at the particles positions again accordingly to the CIC scheme. If a particle lies close to the interface between two regions at different levels ℓ and $\ell - 1$, the force field is interpolated from the coarsest $\ell - 1$ level. As such, it implies that a particle must penetrate significantly within high resolution regions to be sensitive to their force field, otherwise they will be driven by a lower resolution description of the potential.

Particles are advanced thanks to a mid-point scheme. The $p + 1$ value of the position and the velocities are computed from their p value as:

$$\mathbf{r}_{p+1} = \mathbf{r}_p + \mathbf{v}_{p+1/2} \Delta t^p \quad (5)$$

where the mid-point velocity is provided by

$$\mathbf{v}_{p+1/2} = \mathbf{v}_p + \mathbf{f}_p \frac{\Delta t^p}{2} \quad (6)$$

and finally corrected as:

$$\mathbf{v}_{p+1} = \mathbf{v}_{p+1/2} + \mathbf{f}_{p+1} \frac{\Delta t^p}{2} \quad (7)$$

2.2.2 Hydrodynamics

Hydrodynamics are solved through a Eulerian description of conserved fluid quantities, which obey the set of Euler equations (see e.g. Toro (1997)):

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y} + \frac{\partial \mathbf{H}(\mathbf{U})}{\partial z} = \mathbf{S}, \quad (8)$$

where $\mathbf{U} = (\rho, \rho u, \rho v, \rho w, E)$ is the array of conserved quantities, the density, the 3 components of momentum and the energy. \mathbf{F} , \mathbf{G} and \mathbf{H} are the 3 flux functions with

$$\mathbf{F}(\mathbf{U}) = (\rho u, \rho u^2 + p, \rho uv, \rho uw, u(E + p)) \quad (9)$$

$$\mathbf{G}(\mathbf{U}) = (\rho v, \rho uv, \rho v^2 + p, \rho vw, v(E + p)) \quad (10)$$

$$\mathbf{H}(\mathbf{U}) = (\rho w, \rho uw, \rho vw, \rho w^2 + p, w(E + p)) \quad (11)$$

In association with conserved quantities, this is usual to also consider the primitive quantities, namely the density ρ , the velocities $\mathbf{v} = (u, v, w)$ and the pressure p . The total energy is given by

$$E = \frac{\rho}{2}(\mathbf{v}^2) + \frac{p}{\gamma - 1}, \quad (12)$$

with contribution of the kinetic and the internal energy. Here γ is the usual adiabatic exponent, equal to $5/3$ for an ideal mono-atomic gas. The High-Mach flow regime is taken in account using the recipe described in Rasera & Teyssier (2006).

The set of Euler equation is solved in a split fashion, dealing first with the pure transport part with a null r.h.s in eq. 8 then updating the solution by adding the contribution of source terms. The transport update of \mathbf{U}^p into \mathbf{U}^{p+1} is done explicitly (in 1D here for simplicity) via:

$$\mathbf{U}_i^{p+1} = \mathbf{U}_i^p + \frac{\Delta t}{\Delta x} (\mathcal{F}_{i-1/2}^p - \mathcal{F}_{i+1/2}^p). \quad (13)$$

This solution requires the knowledge of intercell fluxes $\mathcal{F}_{i\pm 1/2}^p$ at instant p between cells i and $i \pm 1/2$, through the resolution of Riemann problems. Here we use a MUSCL scheme coupled to an HLLC Riemann Solver (see e.g. Toro et al. (1994); Toro (1997)). First, conserved quantities are linearly reconstructed at the cell boundaries (in 1D for simplicity) :

$$\mathbf{U}_i^{L/R} = \mathbf{U}_i^p \pm \frac{\Delta_i}{2}. \quad (14)$$

Here L/R designates the left and right reconstructed states at the cell i boundaries in $x = 0$ and $x = \Delta x$, the cell center being in $\Delta x/2$. Δ_i is the slope vector of the conserved quantities and is computed using neighbor cells values and a MinMod limiter to ensure a monotonic solution. Before solving the Riemann problem, the boundary extrapolated values are also evolved by a time $\Delta t/2$:

$$\bar{\mathbf{U}}_i^{L/R} = \mathbf{U}_i^{L/R} + \frac{\Delta t}{2\Delta x} [\mathbf{F}(\mathbf{U}_i^L) - \mathbf{F}(\mathbf{U}_i^R)]. \quad (15)$$

The Riemann problem at intercell positions are solved using these evolved states, for instance $\mathcal{F}_{i+1/2}^p$ is obtained from states $\bar{\mathbf{U}}_i^R$ and $\bar{\mathbf{U}}_{i+1}^L$. The MUSCL scheme achieves second order accuracy (Toro et al. 1994; Toro 1997). As discussed previously, when dealing with cells at the interface between 2 regions with different resolutions, coarse quantities are interpolated in a conservative manner at the locations of the ‘virtual’ fine cells. At such interfaces, the low resolution values are assumed to be constant in time and accuracy reduces to first order (Khokhlov 1998; Teyssier 2002). The

multi-dimensionality of the problem is dealt through an unsplit approach with a flux contribution of the 3 directions included at once in the update.

The stability of the solution is obtained by enforcing the Courant condition on the time step

$$\Delta t = C \frac{\Delta x}{3V_h}, \quad (16)$$

where V_h is an estimate of the largest wave speed present throughout the computational domain and $C < 1$. We follow the simple suggestion of Toro with

$$V_h = \max\{|\mathbf{v}|_i + a_i\} \quad (17)$$

where a_i stands for the sound speed at location i . It usually overestimates the limiting velocity and therefore underestimates the corresponding time step, but provides a robust estimation.

The source term in Eq. 8, models deviation to pure conservation. In our case \mathbf{S} holds the contribution of the gravitational force to momentum and energy variation and expresses the coupling of gravitation with hydrodynamics :

$$\mathbf{S} = (0, -\rho \vec{\nabla} \phi, -\rho \mathbf{v} \vec{\nabla} \phi), \quad (18)$$

where ϕ is the gravitational potential. Its inclusion in the solution (eq. 13) is done in a 'explicit' manner with

$$\mathbf{S} = (0, -\rho^p \vec{\nabla} \phi^p, -\rho^p \mathbf{v}^p \vec{\nabla} \phi^p). \quad (19)$$

This contribution modifies the conservative quantities after the pure transport update. However they are taken in account within the MUSCL scheme to compute the interpolated left/right states before solving the Riemann problems. Of course hydrodynamics is also coupled to radiative physics and thermo-chemistry via the internal energy (or equivalently the pressure) of the gas : photo-heating and cooling act as source and sink terms of this quantity and are treated by the radiative transfer engine. Finally, the hydro-engine is able to handle passive scalars that are being advected with the fluid. Among such scalars, one can currently find the hydrogen neutral fraction or in forthcoming developments one can think of the metallicity.

2.2.3 Radiative Transfer

Propagation of radiation is dealt with using a moment based description : light is described as fluid, where its phase space distribution is averaged on velocities to focus on spatial fields. Among this family of radiation description, **EMMA** relies on the M1 approximation (Levermore 1984; González et al. 2007; Aubert & Teyssier 2008).

Taking the first two moments of the Liouville Equation leads to the conservation equations of the ionizing photons density $N_\nu(\mathbf{r}, t)$ and flux $\mathbf{F}_\nu(\mathbf{r}, t)$, in each bin of frequency ν :

$$\frac{\partial N_\nu}{\partial t} + \frac{\partial \mathbf{F}_\nu}{\partial \mathbf{r}} = S_\nu - \kappa_N N_\nu, \quad (20)$$

$$\frac{\partial \mathbf{F}_\nu}{\partial t} + c^2 \frac{\partial \mathbf{P}_\nu}{\partial \mathbf{r}} = -\kappa_F \mathbf{F}_\nu. \quad (21)$$

Here, $\mathbf{P}_\nu(\mathbf{r}, t)$ stands for the radiative pressure tensor. This system of equation being opened, a closure relation is required to be able to solve it. The M1 closure relation is

given by:

$$\mathbf{P}_\nu = \mathbf{D}_\nu N_\nu \quad (22)$$

$$\mathbf{D}_\nu = \frac{3\chi - 1}{2} \mathbf{I} + \frac{1 - \chi}{2} \mathbf{n} \times \mathbf{n}. \quad (23)$$

Here, \mathbf{D}_ν stands for the Eddington tensor and its expression is set by the value of the χ quantity that varies between 1/3 for a diffusive regime (i.e. $F_\nu \ll cN_\nu$) and 1 for a pure transport regime (i.e. $F_\nu \sim cN_\nu$).

The quantities on the r.h.s. of the conservation equations are the source of photons $S_\nu(\mathbf{r}, t)$ (expressed in photons per unit time per unit volume) and the two absorption terms, driven by the absorption coefficients κ_N and κ_F , considered equal in **EMMA** with $\kappa_N = c\sigma_\nu n_H$. This terms couple the hydrodynamics and the radiative transfer by means of gas absorption with n_H being the density number of neutral gas and σ_ν the photo-ionization cross-section at frequency ν .

In the case of multi-frequency transfer, photons are gathered in so-called 'groups' of frequencies and the flux and number densities of a given group satisfy the above conservative equations. In practice, Eqs 20 and 21 can be integrated between two frequencies:

$$\frac{\partial N}{\partial t} + \frac{\partial \mathbf{F}}{\partial \mathbf{r}} = S - c\sigma_N n_H N, \quad (24)$$

$$\frac{\partial \mathbf{F}}{\partial t} + c^2 \frac{\partial \mathbf{P}}{\partial \mathbf{r}} = -c\sigma_N n_H \mathbf{F}, \quad (25)$$

with

$$N = \int_{\nu_1}^{\nu_2} N_\nu d\nu \quad (26)$$

$$\mathbf{F} = \int_{\nu_1}^{\nu_2} \mathbf{F}_\nu d\nu \quad (27)$$

$$S = \int_{\nu_1}^{\nu_2} S_\nu d\nu \quad (28)$$

$$\sigma_N = \frac{1}{N} \int_{\nu_1}^{\nu_2} \sigma_\nu N_\nu d\nu. \quad (29)$$

Typical frequency groups have limits set by the ionization levels of hydrogen and helium (even though **EMMA** does not currently handle Helium chemistry) i.e. [13.6, 24.6, 54.4] eV or chosen to represent broad classes of different types of radiation such as UV, X and hard X-rays.

In practice, the update of radiative quantities is a two stages process : first a conservative transport is performed, then non-conservative contributions (source and sinks) are added within a subsequent thermo-chemical solver (see Sec. 2.2.4). The set of homogeneous (with zero r.h.s) coupled equation is solved for $\mathbf{U} = (N, \mathbf{F})$ using fluxes $\tilde{\mathbf{F}} = (\mathbf{F}, c^2 \mathbf{P})$ with a simple explicit finite difference scheme (in 1D for sake of simplicity):

$$\mathbf{U}_i^{p+1} = \mathbf{U}_i^p + \frac{\Delta t}{\Delta x} (\mathcal{F}_{i-1/2}^p - \mathcal{F}_{i+1/2}^p) \quad (30)$$

that must satisfies the usual Courant Condition:

$$c \leq \frac{\Delta x}{\Delta t}. \quad (31)$$

Here $\mathcal{F}_{i-1/2}^p(\mathbf{U})$ represents the flux function at instant p measured at the interface between the cell i and $i - 1$. This intercell flux is obtained by solving a typical Riemann prob-

lem at this interface : in **EMMA**, this flux is given by the Lax-Friedrich Formula:

$$\mathcal{F}_{i+1/2}(\mathbf{U}) = \frac{\tilde{F}_i + \tilde{F}_{i+1}}{2} - c \frac{\mathbf{U}_{i+1} - \mathbf{U}_i}{2}. \quad (32)$$

As in **ATON**, an implementation of the less diffusive HLL flux is also on the way. At this stage, conservative transport is completed and radiative quantities (N, \mathbf{F}) are in an intermediate state, waiting for the contribution of source and sinks to be taken in account, as explained in the next section.

Before, it should be noted that the Courant condition ensures the stability of the scheme but impose that the numerical 'sampling velocity' $\Delta x/\Delta t$ must be greater than the speed of light. The cost of simplicity provided by the explicit solver is therefore a very fine temporal sampling, greatly enhancing the CPU-cost of the radiative transfer. This stringent constraints on the time step imposed by the Courant condition can be circumvented in two ways. The first one is simply by taking advantage of hardware acceleration (as in **ATON**). In **EMMA** this route is explored with GPUs as accelerating devices and described in section 4. The second one is to reduce the speed of light (Gnedin & Abel 2001; Rosdahl et al. 2013), taking advantage of the fact that in a large number of situations, the effective propagation of radiative information is performed through ionization fronts that propagate at smaller pace. This option can also be set in **EMMA**, as done for instance in Sec. 6.6.1.

2.2.4 Thermal and chemical processes

The thermal and 'chemical' processes encompass the atomic physics that will affect hydro and radiative quantities. Currently **EMMA** only handles atomic hydrogen processes. They contribute to change the number density and flux of photons (source and sinks), the ionization state of the gas and finally its internal energy:

$$\frac{dN}{dt} = S - c\sigma_N n_H N + (\alpha_A(T) - \alpha_B(T))x^2 n_0^2, \quad (33)$$

$$\frac{d\mathbf{F}}{dt} = -c\sigma_N n_H \mathbf{F}, \quad (34)$$

$$\frac{dn_H}{dt} = (\alpha_A(T)x^2 - \beta(T)x(1-x))n_0^2 - cn_H\sigma_N N, \quad (35)$$

$$\frac{de}{dt} = cn_H\Sigma_E N - \Lambda(n_0, x, T), \quad (36)$$

where $n_H = (1-x)\rho/m_p = (1-x)n_0$ is the number density of hydrogen atoms (with individual mass m_p), x being their ionized fraction and n_0 being the total number of protons (i.e. neutral + ionized hydrogen). The quantities $\alpha_{A/B}$ and β are respectively the case A/B recombination and collisional ionization rates. Eqs. 33 and 34 describe the influence of source and sinks on the radiative quantities : in particular the last term in Eq. 33 refers to the recombining ionizing radiation. Note that the on-the-spot approximation can easily be applied by setting $\alpha_A = \alpha_B$ in Eqs. 33 and 35. Eq. 35 details the competing effects of recombination and ionizations on the number of neutral hydrogen atoms. Eq. 36 encodes the evolution of the internal energy density of the gas $e = p/(\gamma - 1)$ due to atomic cooling (given by the cooling rate Λ) and to the photo-heating above the ionization threshold $\mathcal{H} = cn_H N \Sigma_E$. Here $\Sigma_E = (\sigma_E \langle E \rangle - \sigma_N E_{13.6})$ and the quantity σ_E is the energy averaged cross-section over the

group of frequencies of interest:

$$\sigma_E = \frac{1}{N \langle E \rangle} \int_{\nu_1}^{\nu_2} \sigma_\nu N_\nu h\nu d\nu, \quad (37)$$

$$\langle E \rangle = \frac{1}{N} \int_{\nu_1}^{\nu_2} N_\nu h\nu d\nu \quad (38)$$

where the latter quantity $\langle E \rangle$ is the average photon energy in the same group.

In the case of multi-frequency transfer, Eqs (33-36) are solved for each frequency interval (i.e. group) $[\nu_i, \nu_{i+1}]$ where i stands for a group label. Within each frequency group i the corresponding photon density and fluxes (N_i, \mathbf{F}_i) satisfy :

$$\frac{dN_i}{dt} = S_i - c\sigma_{N,i} n_H N_i + (\alpha_A(T) - \alpha_B(T))\delta_{i,1} x^2 n_0^2 \quad (39)$$

$$\frac{d\mathbf{F}_i}{dt} = -c\sigma_{N,i} n_H \mathbf{F}_i. \quad (40)$$

These equations depend on the group cross-section $\sigma_{N,i}$ and source function S_i obtained using Eqs 28 and 29 on the $[\nu_i, \nu_{i+1}]$ interval. The number of radiative conservative updates therefore scales as the number of frequency groups. The Kronecker symbol $\delta_{i,1}$ in Eq. 39 implies that the recombining radiation only contributes to the first group of ionizing photons, the closest to the hydrogen ionization frequency (see e.g. Rosdahl et al. (2013)). Again, if the on-the-spot approximation is used, this contribution is set to zero for all the groups. The thermo-chemical equations Eqs. 35 and 36 are also modified and the following expressions must be replaced:

$$cn_H\sigma_N N \rightarrow cn_H \sum_{i=1}^{N_{\text{groups}}} \sigma_{N,i} N_i, \quad (41)$$

$$cn_H\Sigma_E N \rightarrow cn_H \sum_{i=1}^{N_{\text{groups}}} \Sigma_{E,i} N_i. \quad (42)$$

N_{groups} stands for the total number of frequency intervals considered. Eqs. 41 and 42 depend on the groups cross-sections and photons densities and effectively couple photons of different frequencies that are otherwise transported without any interactions.

In practice, we solve the equations in the same order as above by sub-cycling the dynamical step Δt with a chemical time step $\Delta\tau$. During the latter, all quantities are updated from state p to $p+1$. $\Delta\tau$ is first evaluated from the p state of the internal energy, with $\delta\tau = e^p/(\mathcal{H}^p - \Lambda^p)$. Then all the quantities are updated in the following manner and in that order:

$$N^{p+1} = \frac{N^p + (S + (\alpha_A(T) - \alpha_B(T))x^2 n_0^2)\Delta\tau}{1 + c\sigma_N(1-x^p)n_0} \quad (43)$$

$$\mathbf{F}^{p+1} = \frac{\mathbf{F}^p}{1 + c\sigma_N(1-x^p)n_0} \quad (44)$$

$$x^{p+1} = 1 - \frac{\alpha_A(T^p)x^p n_0 \Delta\tau + 1 - x^p}{1 + \Delta\tau(\beta(T^p)x^p n_0 + c\sigma_N N^{p+1})} \quad (45)$$

$$e^{p+1} = e^p \quad (46)$$

$$+ \Delta\tau c(1-x^{p+1})n_0 N^{p+1} \Sigma_E \quad (47)$$

$$- \Delta\tau \Lambda(n_0, x^{p+1}, T^p)). \quad (48)$$

In this sequence, new information on a given quantity is immediately used to compute a subsequent one. We follow

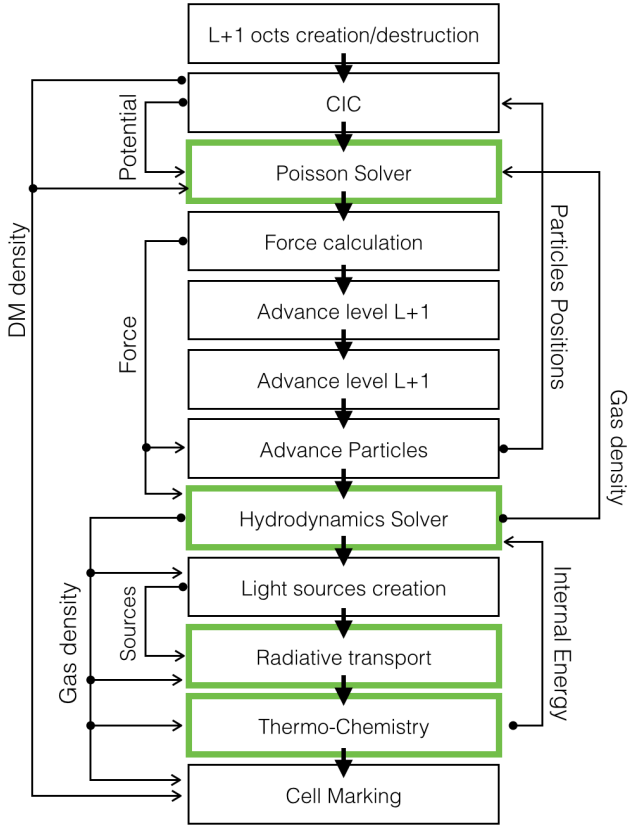


Figure 1. Flow of sequence for a time step at a given level (from top to bottom). Side arrows describe the exchange of physical quantities between different modules to emphasize the most important couplings. Green boxes stand for modules that have been ported on GPU using CUDA. It is assumed that $\Delta t_\ell = 2\Delta t_{\ell+1}$ and the $\ell + 1$ level is advanced twice. This sequence is repeated recursively for all the finer levels.

Rosdahl et al. (2013) and enforce that the internal energy must at most vary by 10% (relatively) during $\Delta\tau$, otherwise the set of equation is recomputed with a time step divided by 2 until this condition is satisfied. In all our experiments, this procedure has been found to be accurate and robust enough. All the rates required to describe the atomic processes such as the recombination, the collisional ionization, and the cooling are taken from the compilation of Theuns et al. (1998). Photo-ionization cross-sections are taken from Hui & Gnedin (1997).

2.3 Time stepping

The organization of time step is intimately constrained by the multi-level structure of the data. A single level time step is organized in quite an usual fashion and described in Fig. 1. Coupling between the different physics occur at different levels, the most explicit ones being :

- between Gravity and Hydrodynamics through the total matter density and the gravitational force it creates,
- between radiation and hydrodynamics through the density distribution, or the gas temperature.

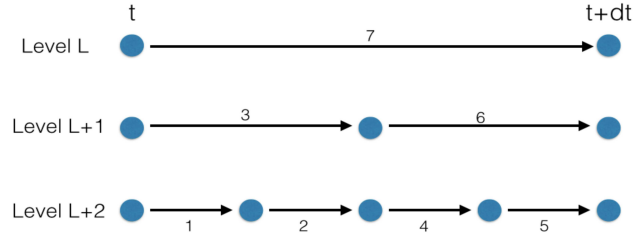


Figure 2. The time-step hierarchy of 3 level of the AMR structure. The coarse level ℓ is advanced from t to $t + dt$, which implies partial advance and temporary updates of fine levels $\ell + 1$ and $\ell + 2$ until they are synchronized with ℓ . Arrows and numbers indicate the sequence of partial updates at different levels to perform this full update of the 3 nested levels.

On top of these explicit coupling, detailed in the next section, implicit ones also occur where e.g. a photo- evaporating halo could in principle affect the underlying dark matter distribution (in the same way supernovae feedback could affect it, even though they are not explicitly coupled, e.g. Pontzen & Governato 2012)

When mesh refinement is enabled, coupling between levels must be taken into account with special care. In EMMA, updates are performed level by level, each level being updated with its own time step Δt_ℓ . Typically, $\Delta t_\ell \sim \Delta t_{\ell+1} \times 2$ (ℓ corresponding to a level coarser than $\ell + 1$). Advancing the solution on a level ℓ can be expressed as the following recursive expression:

$$A_\ell = R_{\ell+1} P_\ell A_{\ell+1} U_\ell M_\ell, \quad (49)$$

where A_ℓ stands for 'advancing the solution' at level ℓ . R_ℓ is the refinement operator (i.e. creating/destroying $\ell + 1$ octs from ℓ cells). P_ℓ and U_ℓ are the operators for the Poisson resolution and the update (i.e. moving particles, updating Eulerian fields) at level ℓ . Finally, M_ℓ corresponds to the marking of the level ℓ cells for future refinements. The recursion is stopped at the maximal allowed level with $R_{\ell_{\max}} = M_{\ell_{\max}} = A_{\ell_{\max}+1} = 1$. For instance, a simulation with three resolution levels (e.g. $\ell = 5, 6, 7$) will be fully updated on Δt_5 according to the following operations (Fig. 2 details this step in a schematic manner) :

$$\begin{aligned} A_5 &= R_5 P_5 [R_6 P_6 [A_7] U_6 M_6] [R_6 P_6 [A_7] U_6 M_6] U_5 M_5 \\ A_7 &= P_7 U_7 P_7 U_7 \end{aligned} \quad (50)$$

where we assumed that $\Delta t_5 = 2\Delta t_6 = 4\Delta t_7$. More generally, the time steps are constrained by :

$$\Delta t_{\ell+1}^1 + \Delta t_{\ell+1}^2 \leq \Delta t_\ell. \quad (51)$$

Once the level $\ell + 1$ has been updated by $\Delta t_{\ell+1}^1 + \Delta t_{\ell+1}^2$, the value Δt_ℓ is updated to synchronize the coarse level on the finer one.

As originally described by Khokhlov (1998), this nested hierarchy of time step has some strong implication on the flux update at interfaces between two levels. When conservative quantities are updated on level $\ell + 1$, adjacent ℓ cells

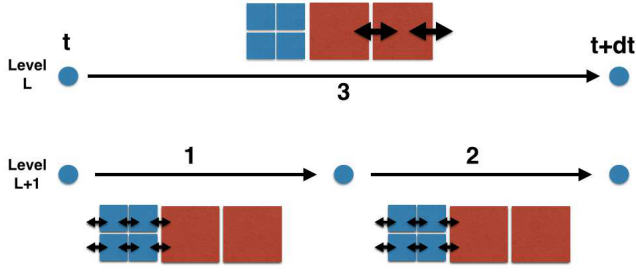


Figure 3. Multilevel update of fluxes at the interface between two levels. To perform an update from t to $t + dt$, fine level $\ell + 1$ must perform 2 steps sub-cycled within a large step of the coarse level ℓ . Updates are performed according to the 1,2,3 sequence : note how the middle coarse cell is updated during the 3 steps with fine fluxes first and a final coarse flux, during the level ℓ update.

must take in account the fluxes produced during the sub-cycling of level $\ell + 1$. For instance, if a conservative quantity U_ℓ and $U_{\ell+1}$ in two adjacent cells must be updated on Δt_ℓ , the following sequence must be obeyed (see also Fig. 3):

- (i) $U_{\ell+1}^p$ is updated to $U_{\ell+1}^{p+1/2}$ using a flux $F(U_{\ell+1}^p, U_\ell^p)$. Meanwhile $U_\ell^{p+1/2}$ is temporarily updated using the same flux.
- (ii) $U_{\ell+1}^{p+1/2}$ is updated to $U_{\ell+1}^{p+1}$ using a flux $F(U_{\ell+1}^{p+1/2}, U_\ell^p)$. Again U_ℓ^{p+1} is temporarily updated using the same flux.
- (iii) Finally, U_ℓ^{p+1} is fully updated taking in account the flux created by adjacent cells of level ℓ .

This partial update of coarse cells during fine time steps is one of the reason to enforce a maximal jump in resolution of unity between two adjacent cells. Larger jumps would create a complex hierarchy of partial update of coarse quantities, difficult to handle properly.

The amplitude of time steps is set by physical temporal scales that must be resolved to track properly their impact on the evolution of quantities. **EMMA** being a multi-physics code, a whole list of time scales are computed for all the cells of a given level and the smallest set its global time step:

- we follow Teyssier (2002), and limit the rate of change of the cosmological expansion factor $\delta a(\Delta t_{\text{cosmo}})/a < \epsilon$,
- a particle cannot move on a scale larger than the size of a cell : $\Delta t_{\text{pic}} = \epsilon \Delta x_\ell / v_{\text{max}}$,
- the local dynamical time must be resolved : $\Delta t_{\text{dyn}} = \epsilon / \sqrt{G\rho}$,
- the hydrodynamical Courant condition must be satisfied : $\Delta t_{\text{hyd}} < \Delta x_\ell / V_h$ (see Eq. 17)
- If light sources are present, the radiation Courant condition must be satisfied : $\Delta t_{\text{rad}} < \Delta x_\ell / c$. It ensures that light propagation is performed in a stable manner.

When full physics are included and effective, the most stringent condition is usually provided by Δt_{rad} and by orders of magnitude since usually $c \gg V_h$. This dominance can be reduced by setting a reduced speed of light. Furthermore as non-linearities increase (and even more when strong shocks such as induced by supernovae feedback will be included) the ratio $\Delta t_{\text{hyd}}/\Delta t_{\text{rad}}$ tends to decrease.

3 COSMOLOGICAL SETTING

Cosmological experiments are implemented using the set of 'super-comoving' variables suggested by Martel & Shapiro (1998). The transformation from physical to supercomoving variables are given by:

$$\tilde{\mathbf{r}} = \frac{\mathbf{r}}{ar_*}, \quad (52)$$

$$\tilde{\mathbf{v}} = \frac{a\mathbf{v}}{v_*}, \quad (53)$$

$$\tilde{\rho} = \frac{\rho}{a^3\rho_*}, \quad (54)$$

$$\tilde{p} = \frac{a^5 p}{p_*}, \quad (55)$$

$$\tilde{\phi} = \frac{a^2 \phi}{\phi_*}, \quad (56)$$

$$\tilde{dt} = \frac{dt}{a^2 t_*}, \quad (57)$$

$$\tilde{N} = a^3 N r_*^3, \quad (58)$$

$$\tilde{\mathbf{F}} = a^4 r_*^2 t_* \mathbf{F}, \quad (59)$$

where starred quantities stand for normalization units with $r_* = L$ (the box length), $t_* = 2/(H_0\sqrt{\Omega_m})$, $v_* = r_*/t_*$, $\rho_* = 3H_0^2\Omega_m/(8\pi G)$ and $p_* = \rho_* v_*^2$. H_0 and $a(t)$ stand for the usual current Hubble parameter and the time-dependent expansion factor.

With this set of transformation, it can be shown that almost all the differential equations to be solved keep their standard expression for a $\gamma = 5/3$ gas. The only notable exception is the Poisson equation which becomes:

$$\tilde{\Delta}\tilde{\phi} = 6a\delta, \quad (60)$$

where $\delta = \tilde{\rho}/\langle\tilde{\rho}\rangle - 1$. Still, this equation remains typical of an elliptic equation that can be solved by all the methods already in place for the Newtonian field equation. Overall, the use of such a transformation greatly simplifies the implementation of cosmological settings in this kind of simulation code.

4 PARALLELIZATION AND VECTORIZATION

EMMA is a parallel code which includes two levels of multi-tasking. The first one is the standard multi-CPU mode, where the computational domain is distributed among several processes that communicate with each other via the MPI protocol. The second level of parallelism resides within an MPI-process where the local load is distributed among several threads of execution. In the case of **EMMA**, this local parallelisation is performed on GPUs but could in principle be extended to other modes of multi-threading such as local shared-memory parallelism among multiple CPU cores or other hardware accelerators. The second level of parallelism can be understood as a vectorization, where arrays of data are processed in parallels through the same set of instructions with minimal communications.

Bearing this two-levels parallelism in mind, **EMMA** has been designed to decouple as much as possible instructions that deal with the logistic of data from the ones that actually perform calculations. Logistics operations are defined as operating directly on the AMR tree : e.g. cell marking

and refinement, tree management and inter-process communications. These operations are handled by CPUs. Computing functions on the other hand expect arrays of data to be 'crunched', without any mention to tree-organized data or inter-process parallelism and return likewise array of results. The physics solvers belong to this second category and are meant to be processed by vector based-hardware, such as multi-core processors, GPUs or any other kind of co-processor. In between, a set of interface functions must be developed to perform gather/scatter operations from/to the AMR tree to/from the calculations arrays. These aspects are developed in the next subsections.

4.1 Distributed parallelism on multiple CPUs

Distributed parallelism is handled through a space-filling curve domain decomposition. Such a curve provides a 1D mapping of a 3D grid by assigning a unique key to each oct as a function of its Cartesian position. The number n_p of parallel processes being defined, the curve is split in n_p successive parts with equal loads, thus assigning a set of octs to each process. The number n_p can thus be arbitrary and the 1D mapping alleviates the need to deal with multiple boundaries along multiple directions. **EMMA** has been implemented with both a Peano-Hilbert space-filling curve and a slab-based key ordering for problems with unidirectional variations (such as the Shock Tube or the Zeldovich Pancake). Currently, the domain decomposition is performed at the level ℓ_c corresponding to the base resolution of the simulation : all the ℓ_c octs are distributed among the processes, in such a way that each process possesses at least one such oct. All octs created from a level ℓ_c cell are assigned to the same process. At the current stage, **EMMA** does not perform any kind of load-balancing that could be obtained by sliding the limits of the 1D domains along the space-filling curve to optimize the distribution of work among the processes.

It should be emphasized that the AMR structure can only be fully exploited if it remains 'consistent': no holes, no level jumps greater than 1 from one cell to another, pointers toward neighboring cells must exist, etc... Furthermore, a given process should be aware, at least partially, of the AMR tree structure of the neighbor processes. **EMMA** copes with these issues by employing the local essential tree decomposition (Warren & Salmon 1993; Dubinski 1996) : each process, even though it has been assigned only a subset of the total volume, is aware of the whole hierarchy of nested octs but only at the levels relevant to its tasks (see Fig. 4). Neighbor cells directly in contact with its domain bring their whole tree from $\ell = 1$ to ℓ_{\max} , those being second order neighbors are one level coarser and so on. In practice, this local hierarchy is obtained at the initial building of the oct-tree: from the root $\ell = 1$ oct, cells are refined down to the coarse level if they belong to the sub-volume assigned to the current processor or if they are direct neighbors of this sub-volume. It produces naturally a local tree for each processor, individually fully consistent and yet aware of the structure of the direct neighbors.

Communications are handled using the MPI-protocol and if a given process P_0 requires data present on other processes, it must be performed explicitly by specifying which octs should come from which other MPI process. This com-

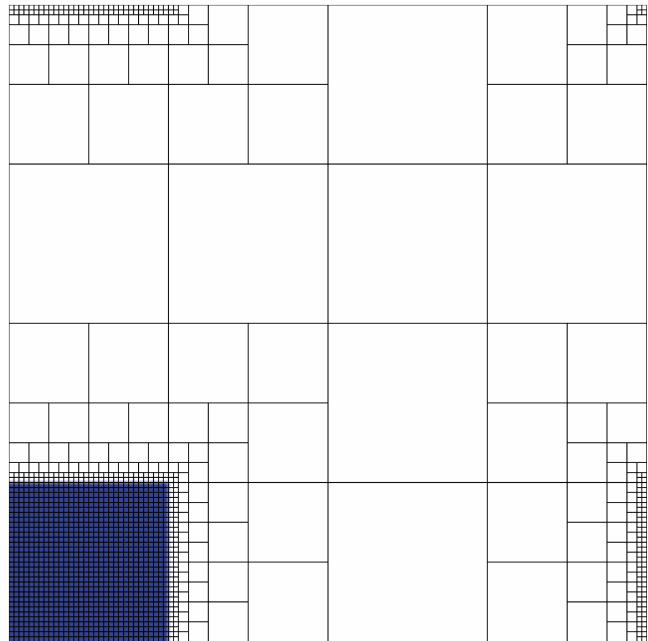


Figure 4. An example of sub-domain seen by a processor in the essential tree paradigm. The processor has been assigned the lower left corner (shaded) but is aware of the whole computational domain. Distant regions are coarsened relative to close ones.

munication protocol in **EMMA** has been written in the following fashion:

- First, all processes P_i build their own lists of neighbor MPI processes $\{P_j\}$ with $i \neq j$.
- For each member of $\{P_j\}$, a list of requests (i.e. of neighbor octs) is established by P_i by storing their space-filling curve keys. This list of keys is sent from each P_i to all its $\{P_j\}$: P_i acts as a client sending requests to neighbor servers.
- Likewise each P_i receives a list of requests from the same sources : P_i acts as a server to neighbor clients.
- Each client key is processed by P_i through an hash table to relate the absolute key to a local pointer to an oct. The data is gathered and sent back to the clients $\{P_j\}$.
- Meanwhile, the data from the servers $\{P_j\}$ is received and scattered back in the local tree by P_i .

This set of instructions is performed level by level and called by the Poisson, the hydrodynamics and the radiation solvers to update border cells that belong to other processes and that have been remotely modified: the flux of information can be considered as outside-in. However there are situation where the information flux is inside-out: a process performed locally will affect directly a value outside its domain. The first example of inside-out communication is the CIC assignment, where a local particles will contribute to an adjacent domain. The second example are the conservative updates due to hydro or radiative fluxes between cells at different levels and belonging to different processes. Because this update is asynchronous between levels, conservative values can be updated in a neighbor coarser cell and must be communicated to its home process. In this case the protocol is similar to the one described above except that there are no request stage and data is sent directly from the server to the client.

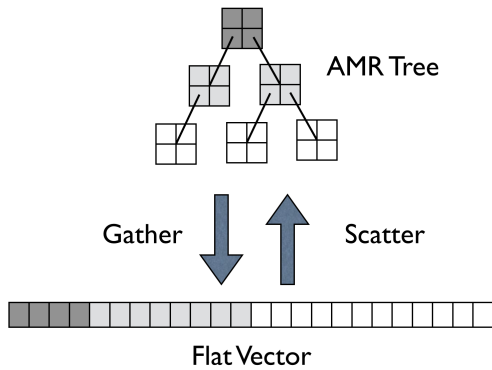


Figure 5. Schematics of the conversion between oct-tree data management by the CPU for the AMR structure and array based calculations, for instance on GPU. A back and forth flow of data is performed through gather/scatter operations. Grey levels label different refinement levels both in the tree and the array.

It should be noted that without load-balancing, the list of neighbors $\{P_j\}$ is static for each P_i . Hence the first stage of the communication protocol has to be performed only once. However the list of *neighbor octs* is dynamic, because of mesh refinement. Therefore, for a given pair client/server P_i/P_j , the list of requests is changing on the fly and should typically be recomputed any time octs have been created or destroyed.

4.2 Local vectorization

Local parallelism relies on a vector-based strategy, where arrays of data are processed through the same set of instructions and possibly on architecture with vectorization capability. The driver for this choice of design is the recent emergence of multi-core processors, graphics processing units or CPU-based co-processors, that all relies on this programming paradigm to be fully efficient. In the case of **EMMA**, this kind of parallelisation is focused on the physics solvers (i.e. the relaxation of the Poisson equation, the conservative update of hydrodynamics and radiative transfer and the chemistry solver) which are fed with arrays of initial states and evolved into arrays of updated values.

Relying on vectors presents several pros. First it guarantees an optimal data layout in general by ensuring that it is accessed in a coalesced and aligned manner : computation directly on tree stored values would induce random and unpredictable memory accesses, whereas an array-based organization ensures proximity of successive or concurrent calculations thus providing optimal performances. For GPUs in particular, enforcing this kind of memory access is a requirement to obtain a maximal throughput of the devices. Furthermore, arrays are generic and simple structures of data, that can be processed in a general manner : each element of an array is computed like an other one and the implementation of this single-element flow of instruction can usually easily ported from one architecture to an another or even from one language to another. The difference usually arises on the details of the scanning operation on all the elements : they can be parsed in sequence for a scalar calculator or by launching multiple threads of a single element computation on GPU or shared memory cores or by taking

advantage of vector abilities of languages such as Fortran 90 or Python. Overall, vectorization provides an opportunity to choose easily a language or an architecture for the code computational modules, without any consideration on the design and layouts of the data structure. For instance, **EMMA** has been coded into both scalar CPU and CUDA GPU versions of the Poisson, hydro and radiative transfer solvers, both versions working in the same AMR framework. In fact, upcoming developments may lead to changes in the way AMR is handled and it would not impact the way physical engines are designed.

However, it becomes readily apparent that the AMR oct-tree being a non-vector based way to store data, the latter must be converted back and forth from a tree-based organization to an array-based one (see Fig. 5). These *gather* and *scatter* operation are critical to the code performance as they constitute bottlenecks to the overall code performances. Nevertheless, if the amount of calculation is large enough, the cost of these operations can be hidden by computing or by overlapping transfers and calculations. In practice in **EMMA**, when data is gathered from the tree to update a value in a given cell, all the necessary values from the neighbors are gathered too. For instance, Eq. 4 requests 7 values to update the potential of a given cell (6 from the cardinal neighbors and 1 for the density). Their related gather operation therefore organizes data in 7 arrays of n_a values, required to update the potential in n_a cells. Of course, since two adjacent cells share some neighbors, the data in these arrays can be redundant. Similarly, intercell fluxes (used during hydrodynamics and radiative transfer) are computed twice for two adjacent cells. In principle such overheads could be avoided but at the cost of coding simplicity and at the current stage the data or flux evaluation has been kept redundant. Gather operations are also in charge of dealing with resolution jumps : if a given cell requests data from a neighbor at an unavailable resolution, it is interpolated linearly from coarse data at the position of the fine virtual cell. Global boundary conditions are also dealt by these gathering operations As said previously, boundary conditions are periodic by nature : if transmissive boundaries are required, the gather operation replaces the data from the periodic neighbor cell by the data of the current cell. If reflective boundaries are set up, the same operation is performed with an additional flux inversion.

5 COARSE RADIATIVE TRANSPORT APPROXIMATION (CRTA)

Sec. 2 describes the standard methodology to couple the different physics within an AMR code with adaptive time stepping. Physical quantities and data structures are updated at the pace of the fastest evolving "dynamics" among the collisionless, the hydro- and the radiative ones. This implementation has been proven to be both accurate and practically sustainable (in terms of required computing resources) for hydro-dynamical codes in the past. However the inclusion of explicit radiative transport and out-of equilibrium chemistry severely impact the code's efficiency as it must track processes on time scales one or two orders of magnitude smaller than the pure hydrodynamical case. Hence an experiment covering a given physical duration must be

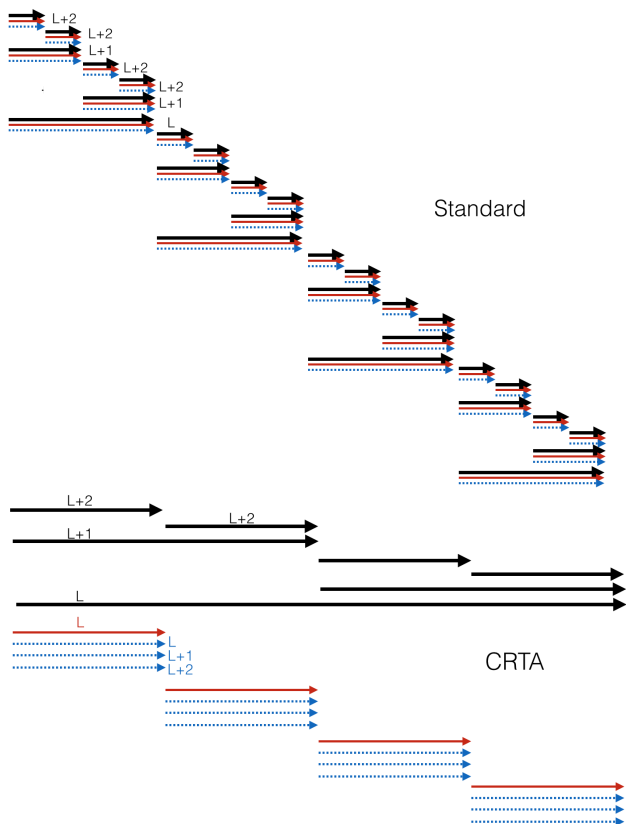


Figure 6. Comparison of the time stepping and multi-physics coupling in the standard AMR approach (top, described in Sec.2) and in the coarse radiative transport approximation (CRTA) description (bottom, described in Sec. 5). Thick black/ thin red/dashed blue arrows stand respectively for dynamics (collisionless+hydro) operations, radiative transport and thermo-chemistry. The thermo-chemistry is shown as a dashed line to indicate that it is already sub-cycled with respect to the radiative transport. In both cases, the simulation is updated by a dynamical time step (from left to right), taken to be equal to 4 times the coarse radiative time step. Note how the red arrows have the same length in both scheme, corresponding to a coarse radiative transport time step. It is also assumed that 2 additional levels of refinement are enabled (L+1 and L+2). In the standard case, such an update takes 4 full updates of the AMR tree, because each update covers a radiative time step, resulting in a total of 83 engine calls. In the CRTA case, the dynamics are updated first over a dynamical time step (7 calls) and radiation + thermo-chemistry are treated in a second stage (16 calls) for a total of 23 calls. Note how the radiation transport is only performed at the coarse level.

sampled with a number of operations one or two orders of magnitude greater, including not only physical engines but also any kind of logistics functions or overhead. As such, any deviation to a perfectly optimized set of operations can see its magnitude multiplied by a factor 100 and reduces significantly the code efficiency.

In this section we suggest an additional level of approximation for the coupling between radiative processes and dynamics (hydro + collisionless). It can significantly reduce the resources necessary for a simulation with radiation, at the cost of a degraded (mostly spatial) resolution. It is summarized in Fig. 6 and relies on two sets of additional ap-

proximations compared to the standard implementation of Sec. 2:

(i) Radiation transport and the associated thermo-chemistry is explicitly decoupled from the dynamics (collisionless+hydro-). Hence to advance the simulation, dynamical quantities are updated first on all the AMR levels on a timescale only constrained by the dynamics (usually set by the hydro- CFL condition). Then, matter is considered as "frozen" and radiation is propagated within this static distribution for the same duration. Since typical speeds encountered in dynamical processes are of the order of the local sound speed or free-fall velocities, which are much smaller than c , such decoupling should remain under control and provide results similar to the standard procedure. Of course, radiation is subject to stringent CFL condition which implies that radiative quantities are updated through an intensive subcycling (typically 100-1000 cycles) of the dynamical time step with small radiative time scales.

(ii) Additionally, radiative transport is only performed at the *coarse* level. However, thermochemistry is still computed on refined levels, but with a coarse-grained description of the radiation field that is simply injected from the coarse to the fine levels. Not only it reduces the number of transport operation but it also reduces significantly the number of thermochemistry steps : this engine is already subcycled even in the standard approach (see Sec. 2.2.4) and can therefore operate on a large radiative time step. Furthermore if an equilibrium situation is encountered in a given cell (a frequent situation in fully ionized regions for instance), this thermo-chemistry subcycling can be reduced to a few cycles.

Fig. 6 provides a simplified comparison of the standard and the coarse radiative transport approximation scheme (CRTA hereafter). We arbitrarily chose a situation where the coarse dynamical time step is 4 times larger than the radiative one at the coarse level. In the standard description the time step is set by the radiative CFL condition at all levels. Hence for a coarse+2 refined levels situation as the one described in Fig. 6 the number of dynamics, radiative transport and thermo-chemistry engine calls are identical and equal to 28 (4 on the coarse level and 24 on the refined ones) for a total of 84 calls. In the CRTA case it reduces to 7 dynamics calls (including 6 calls on refined levels) + 4 radiative transport calls at the coarse level + 12 thermo-chemistry calls for a total 23 engine calls, i.e. a factor of 3 smaller in the number of operations. Bear in mind that a realistic case rather involves a ratio of 100 to 1000 between dynamics and radiative time steps and 5-10 refinement levels: in such cases, the CRTA approximation essentially reduces the cost of hydrodynamics to zero and by neglecting transport on refined levels, it reduces the cost of a radiative time step by a few tens. Additionally, AMR logistics, communications setups, analytics, etc.. are performed only at the dynamical time step and their costs are also essentially set to zero in the CRTA approach. Incidentally, this technique also increases the relative weight of physical engines over numerical overheads and among the engines it increases drastically the weight of radiative transport+chemistry over the others : it turns out that the latter engine is one of the most efficient in the use of vectorization (see Sec. 4) and therefore in the use of hardware accelerators such as GPUs. CRTA is expected to take a greater benefit of such devices to accelerate the code.

Of course this increased efficiency comes at a cost. The most evident one is the decreased spatial resolution for radiative transport, even though the thermo-chemistry is performed at the highest resolution available. It could be thought of as an intermediate approach between an homogeneous radiation field (as usually assumed in non-RT cosmological simulations) and a full AMR description. In the CRTA approach, spatial UV field fluctuations are existent but coarsened. However the impact could be limited. First, as shown in Aubert & Teyssier (2010), radiation fields (radiative density and flux) do not exhibit significant clumping factors compared to the ones of the matter distribution and are relatively smooth even in highly resolved simulations. In fact, the values of radiative densities span orders of magnitude between sources and dark voids and are therefore not very sensitive to local fluctuations. Furthermore, we use here a highly diffusive scheme (based on a LF intercell flux evaluation) which accentuates further the smooth aspect of radiative fields. One could therefore argue that having a fine spatial description of radiative density fields is not absolutely necessary. The other level of approximation is the imperfect coupling of radiation and matter, the latter being considered as still when light is being cast. Somehow, it relates to previous post-processing techniques but performed on the fly at every dynamical time step. Post-processing is known to provide satisfying results for large scale experiments on $\sim 50+$ Mpc scales, hence we can be confident that this imperfect coupling can be controlled in such cases. On the other hand, it is clear that some coupling between matter and radiation in highly refined cells will be lost and it is difficult to evaluate properly this loss. As shown in Sec. 6 the CRTA returns very satisfying results for the tests shown here but in general using CRTA would require to perform an additional level of convergence study to ensure that this imperfect coupling is under control.

6 CODE VALIDATION

EMMA has been submitted to a series of test to validate its implementation. For various combination of simulated physics, documented experiments are described here, as well as the code results.

6.1 1D Hydrodynamics : Shock Tube

The shock tube is a 1D test where a Riemann problem is evolved by means of a simulation. It focuses on implementation of hydrodynamics and the ability of the MUSCL scheme and HLLC Riemann solvers to capture shock features. The initial conditions consist in a jump at $X_0 = 0.3125$ between two different states ($\rho_1 = 1, u_1 = 0, p_1 = 1$ and $\rho_2 = 0.125, u_2 = 0, p_2 = 0.1$, taken from Toro (1997)). The solution to this Riemann problem is known and can therefore be compared to the results delivered by EMMA.

The calculation has been performed using a $\ell = 6$ coarse resolution with 4 additional levels of resolution, triggered by density gradients satisfying $\Delta\rho/\rho \geq 0.015$. Even though the problem is 1D, the calculation has been performed in 3D with the jump occurring along the x direction. Transmissive boundary conditions were retained along the x direction and periodic ones along the two others.

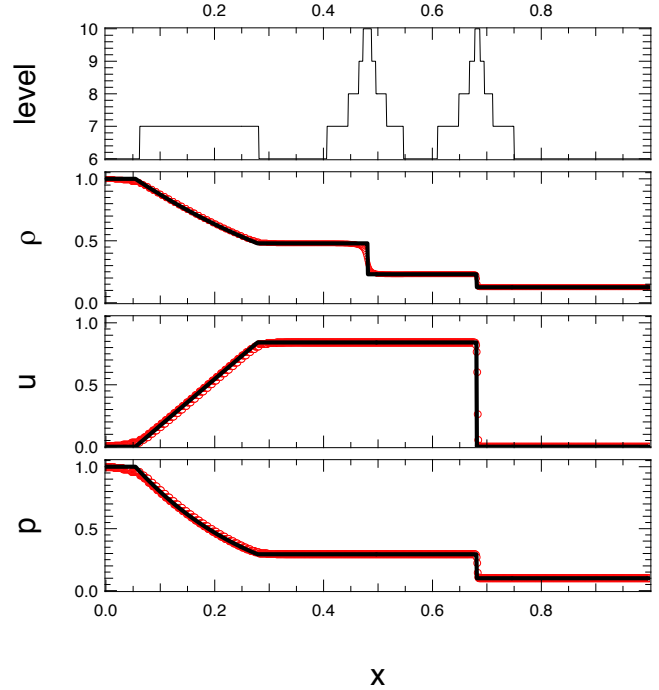


Figure 7. Shock tube experiment. From top to bottom: refinement level, density, velocity and pressure as a function of position. Points stand for the simulation results and solid lines for analytic profiles.

Fig. 7 shows the density ρ , the velocity along the x -direction u , the pressure p and the refinement level at $t = 0.2$. Also shown is the solution of the Riemann problem, as a solid line. Clearly, the match is satisfying with shocks being resolved on a few cells, thanks to both the shock-capturing scheme and the improved resolution allowed by the on-the fly refinement. It can also be noted that the contact wave is also present near $x=0.5$, even though some smearing can still be present at this resolution. Overall, this standard test demonstrates the ability of EMMA to solve classic hydrodynamical problems at high resolution.

6.2 1D Gravity+ Hydrodynamics: Zeldovich Pancake

The Zeldovich Pancake test (Zel'dovich 1970)) tracks the evolution of a single planar mode in an $\Omega_m = 1$ expanding Universe, where the linear stages of the evolution can be analytically predicted. The initial matter density is given by :

$$\rho(x) = 1 + \frac{1+z_c}{1+z_i} \cos(2\pi x) \quad (61)$$

whereas the initial velocity is given by:

$$u(x) = \frac{1}{\pi} \frac{1+z_c}{(1+z_i)^{3/2}} \sin(2\pi x). \quad (62)$$

Here the mode oscillates along the x direction. z_i and z_c stand for the initial and collapse redshift.

As in Sec. 6.1, this test case has been simulated with EMMA in 3D as a planar experiment. Both the baryons and dark matter were included with $\Omega_b = 0.1$. The base resolution is $\ell = 6$ (i.e. 64^3 cells) and the dark matter field

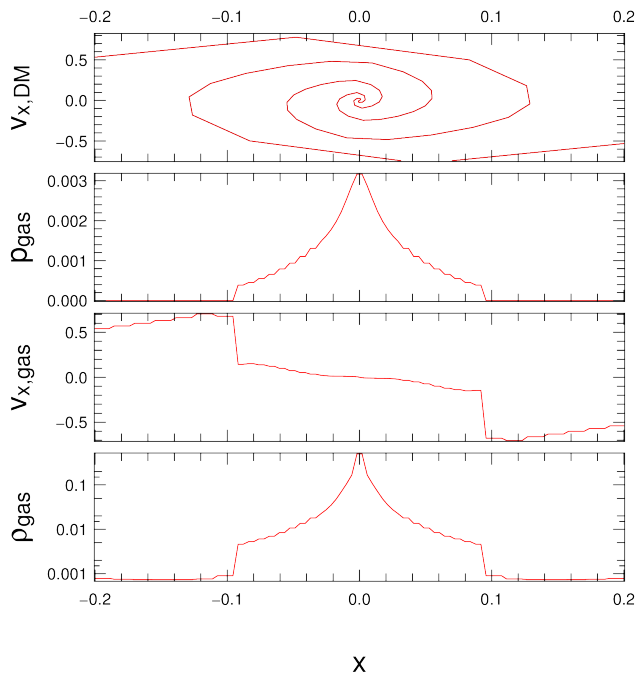


Figure 8. Zeldovich Pancake experiments with baryons ($\Omega = 1, \Omega_b = 0.01$). From top to bottom: the dark matter velocity $v_{x,DM}$, the gas pressure p_{gas} , velocity $v_{x,gas}$, and density ρ_{gas} . All the quantities are shown as a function of the position, at $z=0$. The sine wave is initiated at $z=100$ and collapsed at $z=10$. One can note the increasing resolution toward the center of the caustic, from $\ell = 6$ to $\ell = 8$.

is sampled with 64^3 particles. The initial temperature is chosen to be arbitrarily small at $T = 10K$ and velocities orthogonal to the x-direction are taken to be zero. The experiment has been conducted down to $z = 0$ with $z_i = 100$ and $z_c = 10$. Two additional refinement levels were triggered on gas density gradients $\Delta\rho/\rho > 0.1$. This setting provides a situation where the cosmological setting and the coupling between dark matter and baryons are tested. Linear stages were compared to the analytic solution and were found to match at better than the % level (not shown here) until the redshift of collapse.

Fig. 8 shows the $z=0$ baryon density, velocity and pressure as well as the dark matter phase diagram. Clearly, being way later than the collapse redshift ($z_c = 10$), it can be noted that several ‘plane-crossing’ occurred with a significant number of foldings for the dark matter (DM hereafter) phase space diagram. Baryons fell in the DM potential, creating shocks and an inner increase of temperature (via the pressure) within the collapsed region. In particular one can note how the infall velocity of the gas is strongly reduced as it enters the collapsed gas. Refinement levels were triggered as expected, providing a better resolution of the density peak and a smoother description of the phase space curve of DM in the innermost regions. Finally, direct comparisons with e.g. Teyssier (2002), shows that the results of EMMA are consistent with other codes, even at this latest stages of the pancake collapse.

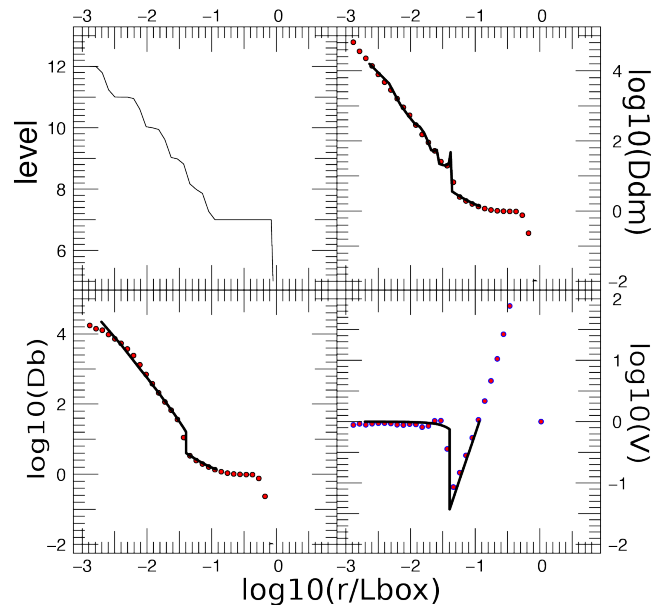


Figure 9. Self Similar 3D Collapse of a Top-Hat density perturbation. Clockwise from top left: the radial profile of refinement level ℓ , dark matter density D_{dm} , baryons radial velocity V and density D_b , at $a=0.56$. As in Bertschinger (1985), these quantities are expressed in units of turnaround quantities. Dots are for simulation results and lines stand for the analytic prediction of Bertschinger (1985). Radii are in units of the box length.

6.3 3D Gravity + Hydrodynamics: Bertschinger’s Self Similar Infall

This experiment aims at reproducing the calculation made by Bertschinger (1985) where a top-hat overdensity within an expanding Einstein-De Sitter Universe ($\Omega_m = 1$) collapses toward a scale-invariant density distribution in a self-similar fashion. Provided that radii are rescaled to the turnaround radius λ , this self-similarity can be fully predicted by analytic means. In the original paper, several configuration are explored and here we focus on the evolution of an overdensity that includes baryons in a dark-matter dominated potential. Compared to the test described in 6.2, the situation here is three-dimensional with a spherical symmetry.

In practice, we generated a regular lattice of DM 128^3 particles starting at $z = 1000$ in a 1 Mpc box, with cosmological parameters $\Omega_m = 1$, $\Omega_b = 0.01$, $H_0 = 70$ km/s/Mpc. Two types of DM particles co-exist : particles within a radius $R_i = 0.05$ Mpc to the center were assigned a greater mass (equals to $7.89 \times 10^4 M_\odot$) than particles at larger distance (with a $6.37 \times 10^4 M_\odot$ mass), in order to produce a central overdensity $\delta_i = 0.2$. Also an arbitrarily cold and motionless gas has been sampled on a coarse grid of 128^3 (i.e. $\ell = 7$), with the same central overdensity as dark matter : cells within a 0.05 Mpc radius were assigned a 797 M_\odot mass and ones at greater radii were given a 604 M_\odot mass. Mesh refinement triggers when the mass within a cell is greater than 8 times the mass of a low-mass DM particle : this criteria is similar to the one used in cosmological simulations in order to provide a quasi-Lagrangian strategy.

Fig. 9 shows the radial profiles of the DM and baryon densities as well as the baryon radial velocity, for the simu-

lation and compared to the fits of the analytic solution provided by Bertschinger (1985). Also shown is the spherical average of the refinement level. As in Bertschinger (1985), the DM density D_{dm} , the baryonic density D_b and the baryonic radial velocity V are expressed in units of ‘turnaround’ values :

$$D_{\text{dm}} = \frac{\rho_{\text{dm}}}{\rho_{\text{ta}}} \quad (63)$$

$$D_b = \frac{\rho_b}{\rho_{b,\text{ta}}} \quad (64)$$

$$V = \frac{v_b}{V_{\text{ta}}}, \quad (65)$$

where $\rho_{\text{ta}} = (6\pi G t^2)^{-1}$ and $\rho_{b,\text{ta}} = \Omega_b \rho_{\text{ta}}$. Bertschinger (1985) gives the evolution of the turnaround radius :

$$r_{\text{ta}}(t) = \left(\frac{4\pi t}{3t_i} \right)^{8/9} \delta_i^{1/3} R_i, \quad (66)$$

where t_i , δ_i , and R_i are respectively the initial time, the initial overdensity and the initial overdensity radius, from which the expression of the associated velocity can be obtained :

$$V_{\text{ta}} = \frac{r_{\text{ta}}}{t}. \quad (67)$$

All the results are shown for $a = 0.56$. Clearly we manage to reproduce the analytic solution in particular the predicted inner logarithmic slope of $-9/4$ for the DM density profile or the shock positions. A small amount of diffusion can be seen in the innermost regions in the baryon density and the velocity jump from the Hubble flow to the shocked region is not as sharp as the predicted one, but the overall features are well reproduced by our calculations and of the same quality described for ART or RAMSES.

6.4 3D radiative hydrodynamics: Growth of an HII region

This test consists of a corner source, powered by a 100 000 K black-body that send photons in a surrounding homogeneous hydrogen-only medium. It belongs to the suite proposed by Iliev et al. (2006, 2009) and comes in two different versions. The first version deals with a static and uncoupled gas (‘Test 2’, Iliev et al. (2006)): we obtained a very good agreement with EMMA (not shown here), which does not come as a surprise since this test was also successfully passed by ATON or RAMSES-RT, that share a great number of details with EMMA. Here we focus on the second version, a coupled test known as ‘Test 5’. The UV photons ionize and heat up the gas, leading to the creation of ionization fronts that propagate inside-out and putting the gas into motion. This test couples radiative transfer and hydrodynamics and has been performed by a whole serie of codes in Iliev et al. (2009).

A 100 000 K Black-body is located at the corner ($x=y=z=0$) of an 15 kpc box, emitting 5×10^{48} UV photons per second. We sample the frequencies with the 3 groups of photons given in Sec 2.2.3. The surrounding gas has an homogeneous number density of 0.001 hydrogen atoms per cm^3 . The calculation is run on a 64^3 coarse grid ($\ell = 6$) and allowing for an additional level of refinement (128^3 i.e. $\ell = 7$) to comply with the resolution requirements of Iliev et al. (2009). The refinement is simply triggered for cells with ionized fraction $0.01 < x < 0.8$: with the two-cells layer of

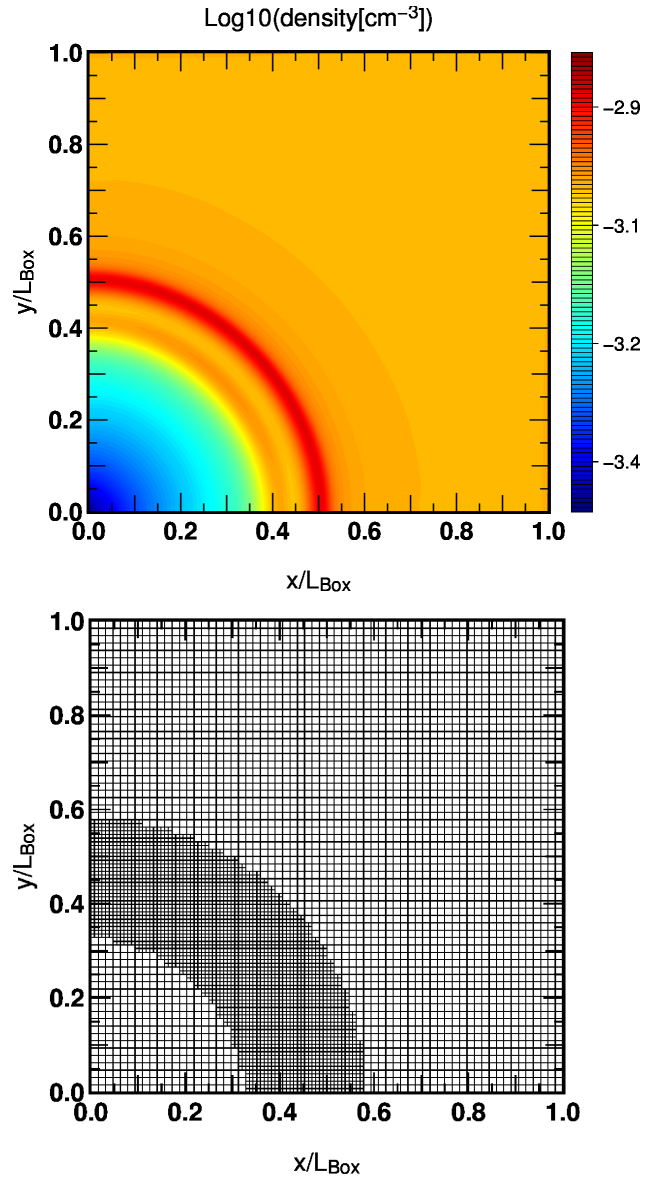


Figure 10. Expansion of an HII region. Top: the log10 of density map (in cm^{-3}), Bottom: the AMR grid used for the computation. The coordinates are expressed in units of the box length which has a physical extent of 15 kpc. The source is located at the bottom left corner and has been ignited 200 Myrs ago.

neighbours being also refined, it provides a simple manner to track the ionization front. Boundary conditions are reflective for boundaries adjacent to the source and transmissive otherwise.

Fig. 10 shows the baryon density in the $z = 0$ plane, as well as the AMR structure that tracks the ionisation front for $t=200$ Myr. The source being in the bottom left corner, the distant cells remain at the base resolution as the front has not yet progressed to these regions. The cells close to the source are also at the base resolution : this region has returned to low resolution as it does not contain an ionization fraction that satisfies the refinement criterion (chosen to track front-like features).

Finally between these two ensembles of coarse cells, one

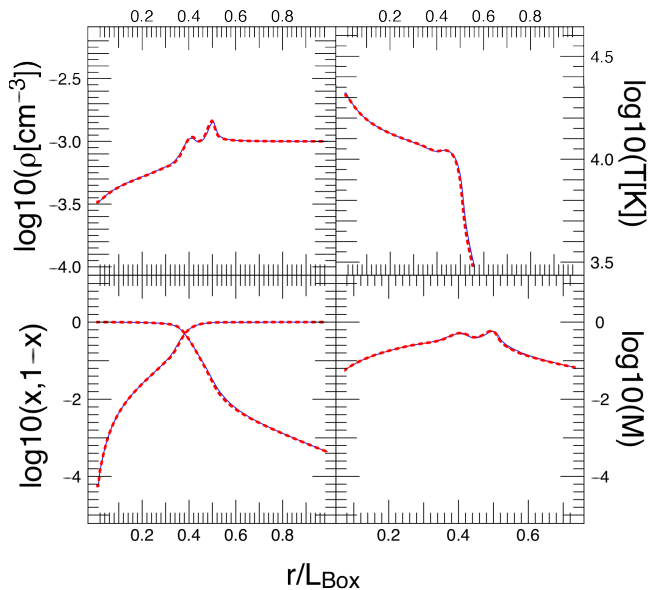


Figure 11. Expansion of an HII region. Clockwise from top left: the gas density, the temperature, the Mach number and the ionized/neutral fraction. The coarse resolution is $\ell = 6$ and refinement is triggered on the ionization front to $\ell = 7$ in accordance with the Iliev et al. test 5. Red dashed lines stand for the radial average taken 200 Myrs after the source has been ignited. Blue lines stand for the same calculation, but performed with the CRTA approximation.

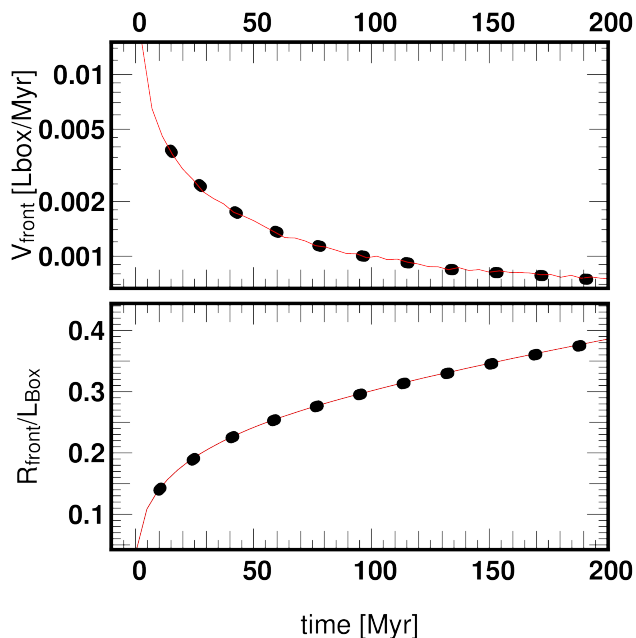


Figure 12. Expansion of an HII region. Front propagation predicted by the standard AMR RT implementation (red line) and the CRTA approximation (dots). Top : the front velocity. Bottom: the front position.

can find the refined region at 128^3 resolution, tracking the front. On the other panel, the number density field of hydrogen is also shown, presenting a typical double peak structure encompassing a void created by the energy injection by the corner source.

Fig. 11 shows the radial profile of the density at the same instant (i.e. 200 Myrs after the source ignition) as well as the temperature, ionisation/neutral fraction and the Mach number profiles. **EMMA** recovers the typical features already obtained by most of the codes of Iliev et al. 2009. The double peak is due to the presence of high energy photons in the spectrum of the 10^5 black body: these photons with large mean free path can deposit energy behind the ionisation front, while lower energy photons deposit their energy at the base of the front. The input of energy at larger radii is also at the origin of the moderate temperature increase seen before its drop in the neutral region. The effect of hard photons can also be seen in the extent of the ionized fraction drop at the front, that would be much sharper in the case of a monochromatic incoming flux. Overall, the fact that **EMMA** reproduces these specific features seen by all other codes indicates that the coupling between radiation and matter and the handling of multi-frequency transfer is consistent with others implementations.

Regarding the CRTA approximation, Fig. 11 and 12 also provides the different fields profiles as well as a comparison of the temporal evolution of the front position and velocity (the front being defined as having a 50% ionised fraction) in both cases. Clearly the CRTA approximation provides the same results as the standard calculation : in Fig. 11 the radial profiles taken at $t = 200$ Myrs are almost indistinguishable and in Fig. 12, the front position and velocities of the CRTA calculation follow the ones obtained from the standard procedure. The evolution is smooth enough both spatially (with features sampled on ~ 15 fine cells or 7 coarse cells) and temporally (with terminal velocities as small as 0.1% box lengths per Myr) to ensure a good convergence of the CRTA toward the standard case.

6.5 3D radiative hydrodynamics: Photo-Evaporation of a dense clump

This situation has also been suggested by Iliev et al. (2009) and consists of a dense cold clump irradiated by a planar UV front. As the ionization front encounters the cloud, the high density will slow down its progression, acting as a trap on the incoming photons. As a side effect, a shadow will also be cast in the trail region of the clump. Finally, the energy deposited in the clump will put the gas in motion, leading to a photo-evaporation process by the incoming photon flux.

The setup is given by Iliev et al. (2009): a spherical clump of radius 0.8 kpc is centered on (5.4, 3.3, 3.3) kpc inside a 6.6 kpc box. Outside the clump, the gas has a 8000 K temperature with a density of 200 atoms/ m^3 . The clump itself has density of 40 000 atoms/ m^3 and a temperature of 100 K. A UV flux with a 100 000K black body spectrum is incoming from the $x=0$ boundary at a rate of 10^{10} photons/ m^2/s . In practice the simulation is performed on a 64^3 grid ($\ell = 6$) with one additional refinement level to comply with the Iliev et al. (2009) recommended resolution. Mesh refinement is triggered for cells with a density greater than the background density. The $x=0$ boundary is a source

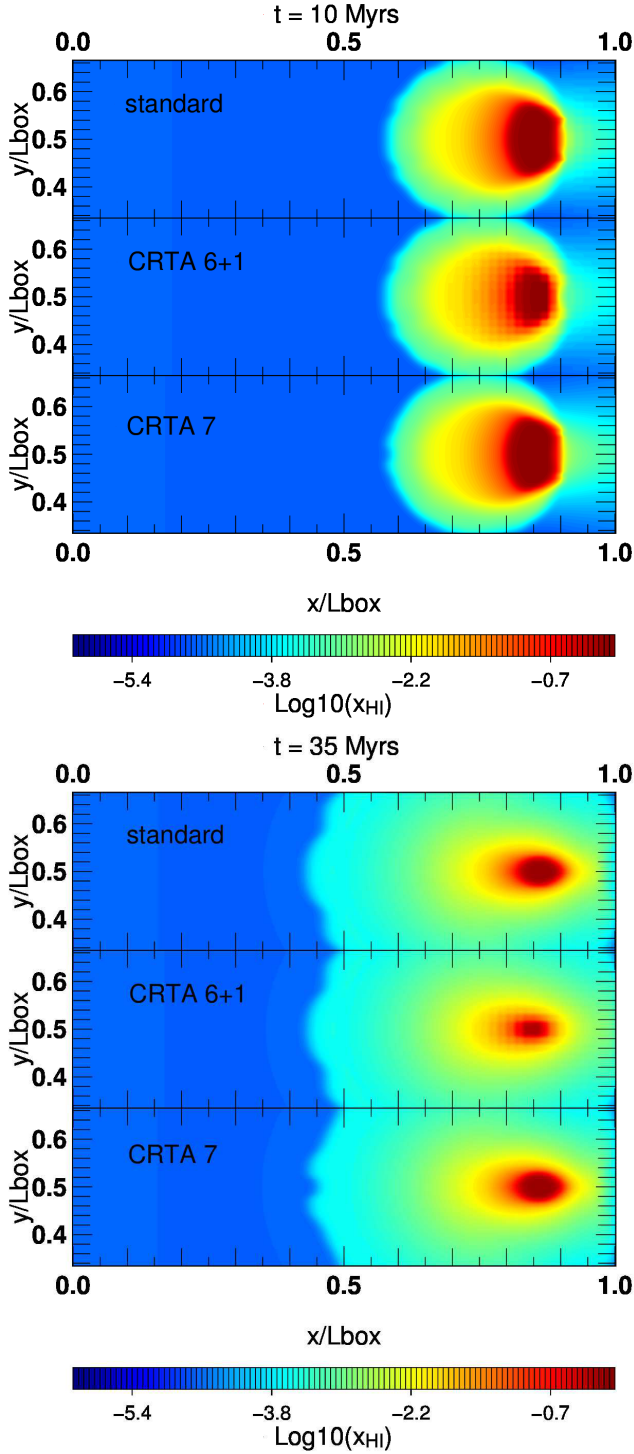


Figure 13. Photo-Evaporation of a dense clump. Top: \log_{10} of the neutral fraction along the plane of symmetry of the clump at $t=10$ Myrs, as predicted by the standard AMR RT implementation (top) and the CRTA approximation (middle), both assuming an $\ell = 6$ base level + 1 level of refinement. The bottom row is the prediction of the CRTA approximation on a static $\ell = 7$ grid. Bottom: the same quantities at $t=35$ Myrs. Blue stands as ionized and red as neutral.

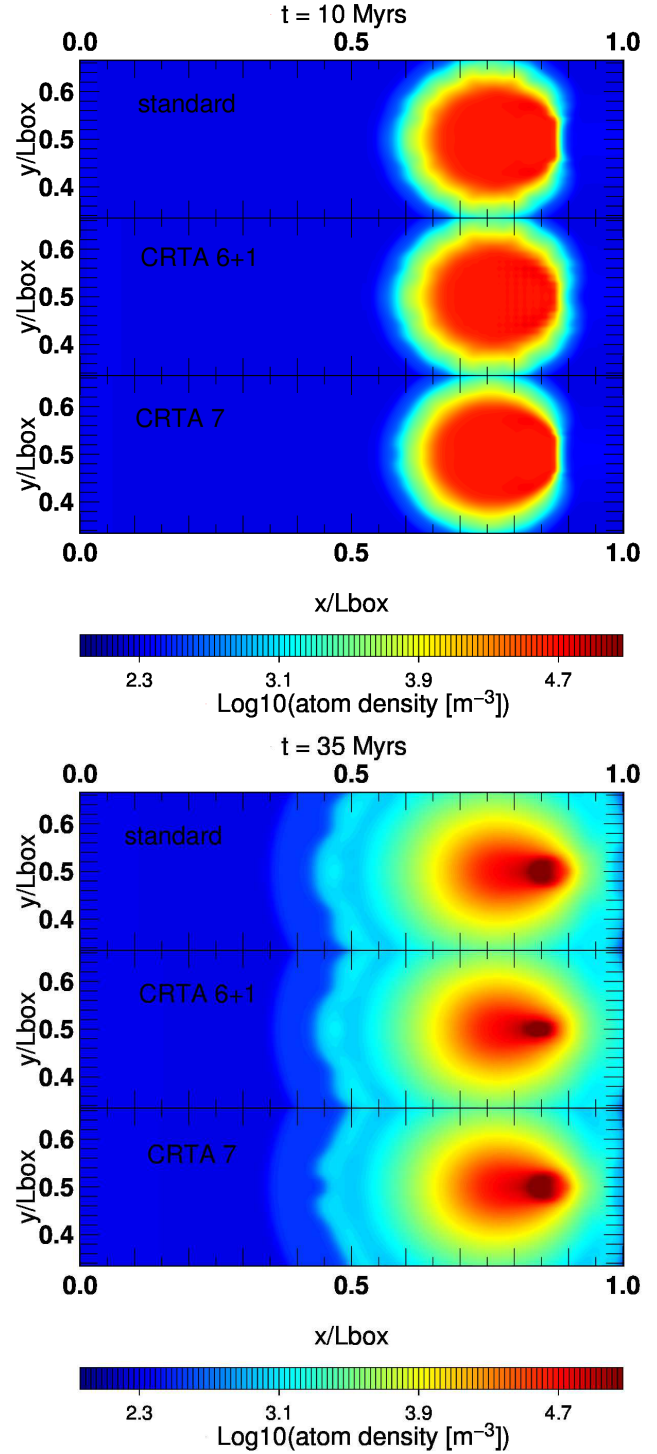


Figure 14. Same as in Fig. 13 but for the gas density. Underdense regions are blue and dense regions are red.

of flux of the required rate, whereas the $x=6.6$ kpc boundary is transmissive. Boundaries in the two other directions are periodic.

Fig. 13 and 14 show maps of the neutral fraction and gas density at $t=10$ and 35 Myrs. In each figure the top and middle row were obtained from 64^3 simulations with an additional level of refinement, the top being obtained from the

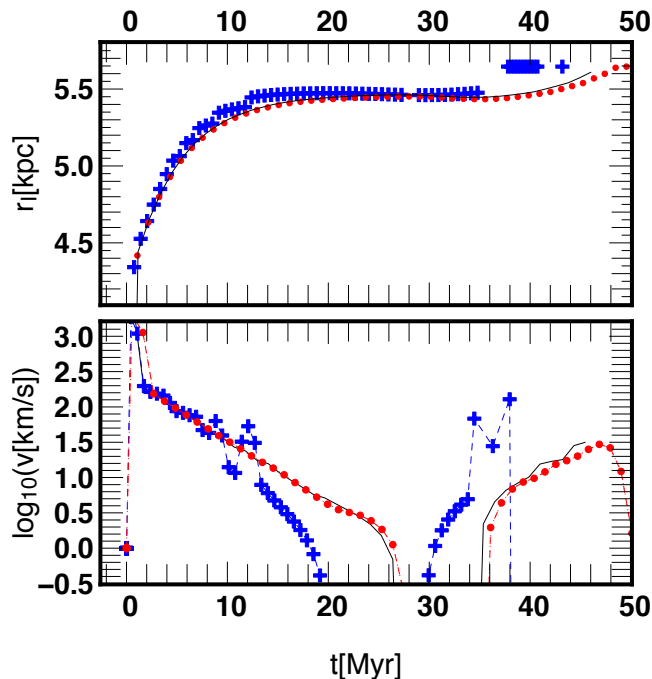


Figure 15. Photo-evaporation of a dense clump. Position (top) and velocity (bottom) of the ionization front along the x axis as a function of time. The solid line stands for the standard AMR RT implementation results and blue crosses stand for the CRTA approximation results. Both were obtained using an $\ell = 6$ base level + 1 level of refinement. Red dots stand for the CRTA approximation results with a $\ell = 7$ static grid.

standard RT implementation on AMR and the middle panel being obtained from CRTA approximation simulations.

Globally a simple comparison to Iliev et al. (2009) results demonstrate their consistency. In particular, the evaporation is made obvious with the expansion of the cloud limits due to the energy injected by the UV front. However clear differences can also be noted : first the shadow behind the clump, albeit existent, is much weaker than in other radiative transfer codes. This does not come as a surprise since EMMA implements the GLF flux to compute intercell exchanges and is known to be very diffusive. It therefore prevents the creation of clear cut shadows as a diffuse component of the flux eats the neutral gas in direction orthogonal to the incoming direction of UV photons. The same effect was already noted for ATON (Aubert & Teyssier 2008). Second, at late time, the contours of the extended cloud are not as spherical as expected and present significant fluctuations around a mean radius. These fluctuations are artifacts of the cloud initial sampling on the coarse $\ell = 6$ grid. The same artifact can be seen e.g. in the FLASH-HC results in Iliev et al. (2009) and also linked to the initial conditions. Comparing the standard RT and the CRTA approximation, it can be seen that the latter provides a faster ionization of the clump. Looking at the Fig. 13, the shadowed neutral clumps are systematically smaller in the CRTA regime. It does not come as a surprise since the RT is performed at the base level only, which increases artificially the extent of the UV flux penetration into the clump and also increases the scheme diffusion. Fig. 15 provides a more quantitative insight on this aspect, describing the front progression in-

side the cold clump and its velocity. The front position is defined as corresponding with the position of the cell having $x = 0.5$. At early stages ($t < 7.5$ Myr), the CRTA and standard description produce an identical front propagation. However it can be noted that the CRTA presents a step-like progression due to the coarser resolution of the radiative transfer which translates into a coarser resolution of the ionized front. Later on, the front is pushed back by the expanding cloud in both descriptions (as can be seen from the receding velocities), but it happens earlier in the CRTA case. Finally the front cannot be tracked for $t > 38$ Myrs, as no cells with a neutral fraction greater than 0.5 can be found anymore. Let us mention that comparisons of the standard calculations with the results presented in Iliev et al. (2009) confirms the capacity of EMMA to track correctly the front propagation within the clump. In particular, EMMA recovers as the other codes the phase where the front is pushed back by the expanding cloud, when $t \sim 35$ Myrs. Globally a faster photo-evaporation of the cloud can be detected in the CRTA approximation compare to the standard AMR description, essentially due to the coarse description of the radiative fields.

Finally, we present in the lower panels of Fig. 13 and 14 the results of a CRTA calculation on a static $\ell = 7$ grid. It allows us to probe the separate the effects of incomplete coupling between dynamics and radiative transfer and of the coarse description of radiation. In this experiment, the CRTA is equivalent to a radiative post-processing of the dynamics but performed on the fly, at the temporal scale of dynamical times and without any impact of a coarsened radiative transfer. Compared to the standard treatment, radiative transfer is subcycled with respect to dynamics, leading to a large number of radiative transfer+thermochemistry calls per single gravity and hydro calculation. Clearly the CRTA greatly reproduces the standard calculation in this case, both in the neutral fraction and density maps and in the propagation of the fronts. It confirms that the coarsened resolution is indeed the reason for an accelerated photo-evaporation of the clump and that the radiation subcycling does not induce significant deviation from the standard treatment.

6.6 Cosmological runs

6.6.1 Preliminary Reionization Simulations

Finally, we present the results of full simulations of cosmological reionization. The focus is put on hydrodynamical simulations with radiative transport and on reionization-related quantities but additional tests on the dark matter haloes mass function or the energy conservation are presented in appendix A. We produced a set of 4 simulations with 4 different specific emissivities for the sources. Each simulation consists in a 4 Mpc/h box sampled with 128^3 base resolution cells and 128^3 dark matter particles. These simulations will be referred as X0.3, X1, X3 and X30. The X1 simulation is a fiducial run with sources emissivities that produce a reasonable ionization history. The three additional cases uses stars with boosted or depleted specific emissivities by the corresponding factor, X0.3 and X30 standing respectively for the dimmest and brightest source model. Initial conditions were produced using Mpggrafic (Prunet

et al. 2008) with a Planck cosmology (Planck Collaboration et al. 2013) ($\Omega_m = 0.315$, $\Omega_\Lambda = 0.685$, $\Omega_b = 0.049$, $n_s = 0.96$, $H_0 = 67$ km/s/Mpc) starting at $z = 80$. Each DM particle weights $4 \cdot 10^6 M_\odot$. AMR is triggered using a quasi-Lagrangian strategy and a cell is refined if it contains more than 8 DM particles. Radiative transfer is run with 3 groups of frequencies ([13.6,24.6] eV, [24.6,54.4] eV and [54.4,1000] eV, dictated by the ionization thresholds of hydrogen and helium).

In addition to hydrodynamics and radiative transfer, we had to implement a simple star formation recipe in order to populate the simulated volume with the ionizing sources that drive the reionization process. This star formation model is briefly described here and will be the subject of a dedicated paper in the near future: its is widely inspired by Katz et al. (1996); Kay et al. (2002); Rasera & Teyssier (2006); Dubois & Teyssier (2008). A cell is said to be prone to star formation if either its gas comoving density (n_*) or its gas density contrast (δ_*) are greater than user-set thresholds. Once a cell is flagged to form stars, the number of stellar particles to be created is drawn from a Poisson law with the λ parameter given by:

$$\lambda = \epsilon \frac{m_{\text{cell}}}{m_\ell} \frac{\Delta t}{t_*}. \quad (68)$$

λ corresponds to the average number of stars created during a time step Δt within a cell that contains a mass of gas given by m_{cell} . The star formation process is controlled by a typical star formation time scale t_* and an efficiency parameter ϵ . The mass of a stellar particle is given by m_ℓ , depends on the level of the cell and is equal to

$$m_\ell = \bar{\rho}_b \delta_* \Delta x^3. \quad (69)$$

The following results were obtained with $\delta_* = 150$, and $t_*/\epsilon = 2\text{Gyrs}$. This values would be considered as 'standard' even though we won't discuss them here and we will explore thoroughly the results dependence on these values in a forthcoming paper. As shown hereafter they nevertheless lead to a star formation and a reionization process in reasonable agreement with constraints. Each stellar particle emits photons for 20 Myrs, with a constant emissivity and assuming a 50 000 K black body spectrum (see Baek et al. (2009)). In practice the source emissivity has been tuned by trial and error to produce a reasonable reionization history, complete at $z \sim 6$ and for the fiducial X1 model, it results in an emissivity of 1.5×10^{16} ionizing photons/sec/stellar kg. Taking the calculation of Baek et al. (2009) as a reference, which assumes a Salpeter IMF and $1 - 100 M_\odot$ mass range, it corresponds to a 15% escape fraction. Again, X0.3, X3 and X30 use emissivities multiplied by the corresponding factor, the X30 model being clearly over-powered and merely used to probe the qualitative behaviour of the code in the regime of strong radiation.

At the current stage we restrict ourselves to this simple model that obviously lacks important ingredients. For instance SN/AGN feedback has not been implemented yet, chemistry is limited to the simple hydrogen and no modification of equation of state is assumed at very high densities. As a consequence the star formation rate is essentially not regulated in this cosmological toy model. Hence the following results could not be considered as definitive regarding

what the code could do but should rather be seen as tests on experimental configurations close to production runs.

Fig. 16 and 17 present the distribution of matter, AMR levels, temperature and hydrogen ionized fraction in a 320 kpc/h thick slab of the X1 run at $z = 6.8$. Clearly, the matter on these scales is already highly structured at $z = 6.8$, with regions having a density contrast greater than 1000. These regions are effectively tracked by the AMR grid and the overall distribution of high resolution grids follow the main features of the filamentary structure in this simulated volume. Sources are created in this overdensities and their radiation leads to large HII regions. One can note how the fronts are locally prevented to progress into the IGM by filaments and dense clumps, leading to complex features in their geometry. It can also be seen that the ionization fronts present a certain extent induced by the larger mean free path of high energy photons. It also leads to a preheating of the gas, behind the ionization fronts, to temperatures close to a few thousands K. Within the ionized regions, a quasi homogeneous temperature close to 10000 K is set by the UV radiation with local fluctuations correlated with the density field. Some shock-heated gas (located at $\sim [2, 1.7]$ Mpc/h) with temperatures greater than 100 000 Kelvins can also be seen.

The fiducial model X1 presents a reasonable reionization history and SFR, in broad agreement with observational constraints (see Fig. 18). Compared to Fan et al. (2006), the ionization happens slightly earlier than observed and correspondingly the photoionization rate at $z \sim 6$ is overestimated when compared to Calverley et al. (2011). The cosmic star formation history is also in excess compared to the observationally deduced rates given by Bouwens et al. (2014). This fiducial model could have benefited from a slightly improved calibration to reproduce the observed data points, however we consider that the current level of agreement is good enough at this stage : let us recall for instance that these simulations lack SN feedback and the small simulated volume could also be inadequate to make quantitative prediction on cosmic averaged quantities. At this stage we merely aim at looking for qualitative and not quantitative clues of the impact of radiation within a cosmological settings.

These clues can be obtained by comparing this fiducial simulation with the 3 other models X0.3, X3 and X30. As expected these models result in different reionization histories that are in place at earlier (resp. later) times for larger (resp. lower) emissivities (see Fig. 18). The SFR however remains essentially unaffected by the change in emissivity, except at later times ($z > 10$) for the models with the brightest sources (X30), leading to a depleted star formation rate.

Fig. 19 presents the baryon fraction and the instantaneous SFR measured in the dark matter halos found at $z = 5.5$ in the different models. At this redshift, the reionization is well advanced in most models except in X0.3 where an 75% ionization level is only achieved. Halos have been detected using the HOP halo finder (Eisenstein & Hu (1998)) and baryons are counted within R_{200} i.e. the radius of the spherical region around each halo with an average density 200 times greater than the cosmic average matter density. ~ 450 halos with a mass greater than $10^8 h^{-1} M_\odot$ (corresponding to 45 particles) are found. Clearly a significant scatter can be found in the distribution of the baryon

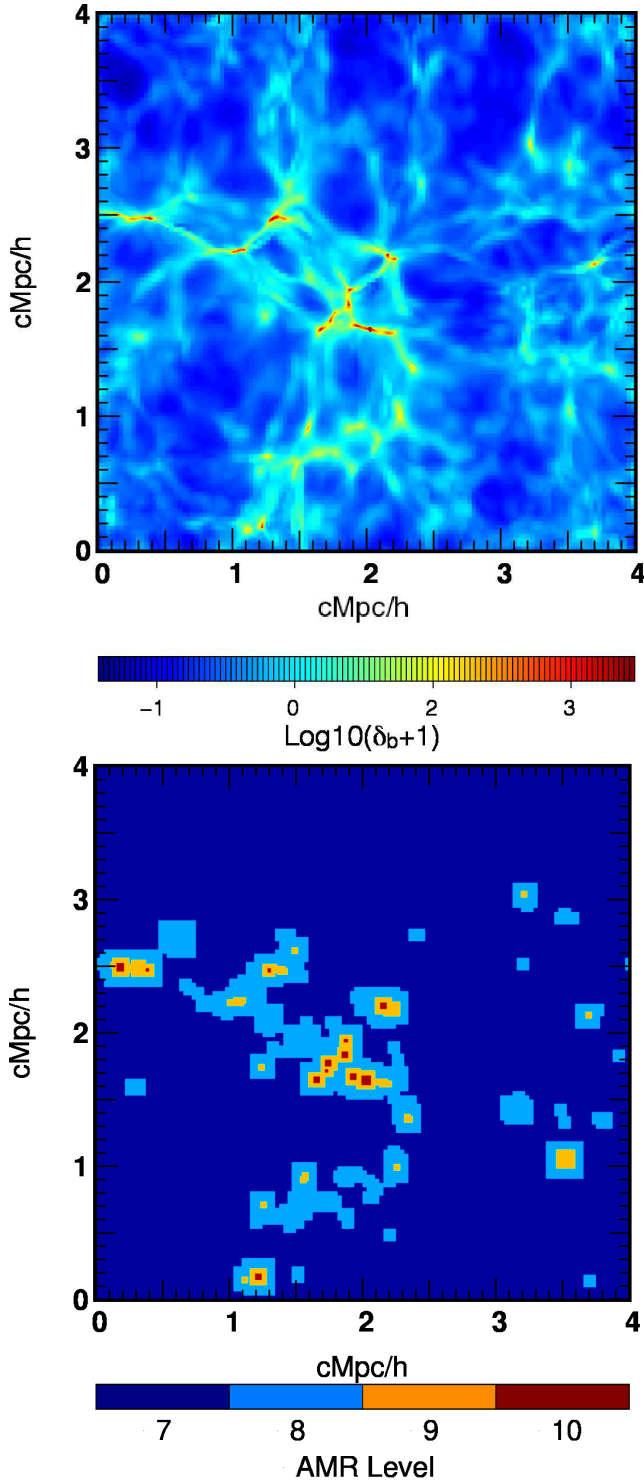


Figure 16. Structuration of matter in a cosmological radiative transfer run of a comoving $4 \text{ Mpc/h} \cdot 128^3$ box, taken at $z = 6.8$. Top: the baryon overdensity map. Bottom: the AMR levels. The shown region has a thickness of 320 comoving kpc/h.

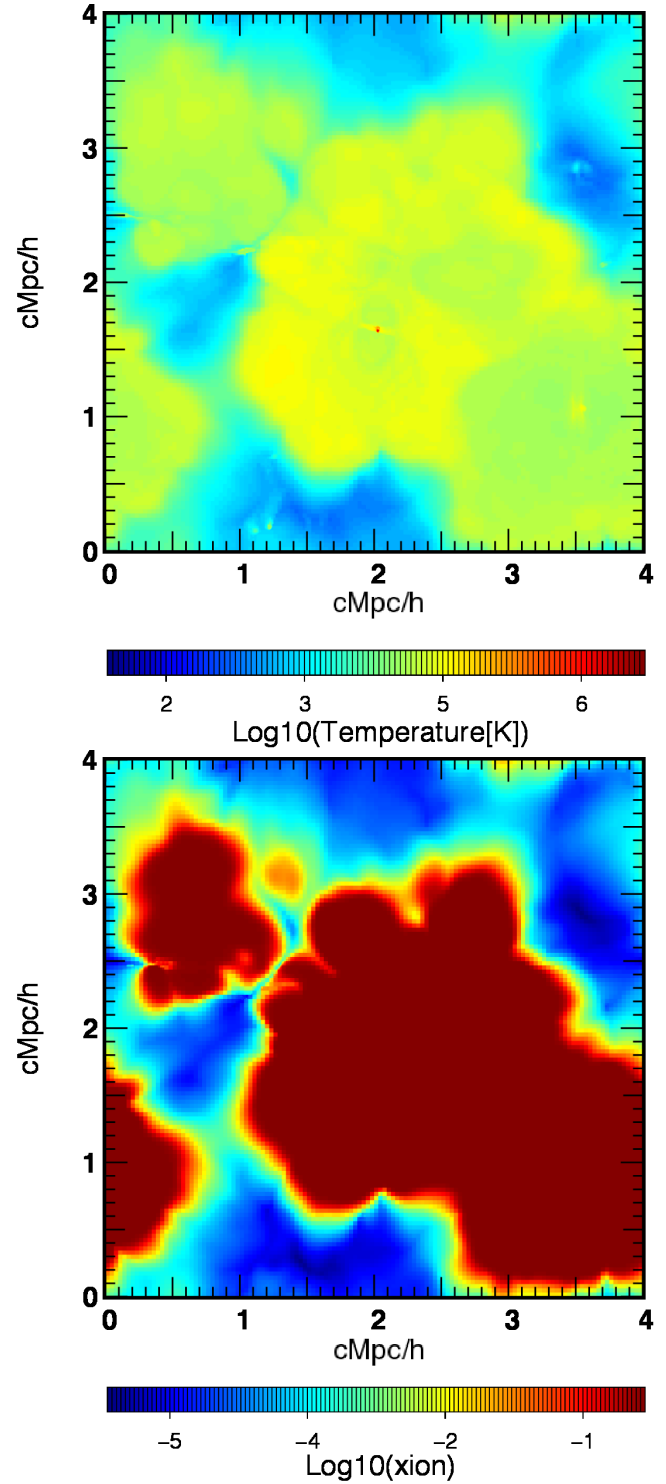


Figure 17. Same experiment and region as in Fig. 16. Top: the temperature map. Bottom: the hydrogen ionized fraction map.

fraction but general trends can nevertheless be observed in the data : halos with a mass greater than $10^9 M_\odot$ basically present a universal fraction whereas lighter objects are more dark matter dominated as expected. A comparison of the fiducial model distribution to the fit provided by Okamoto et al. (2008) shows a reasonable agreement with a correct transition mass at $M \sim 3 \cdot 10^8 M_\odot$, even though a signif-

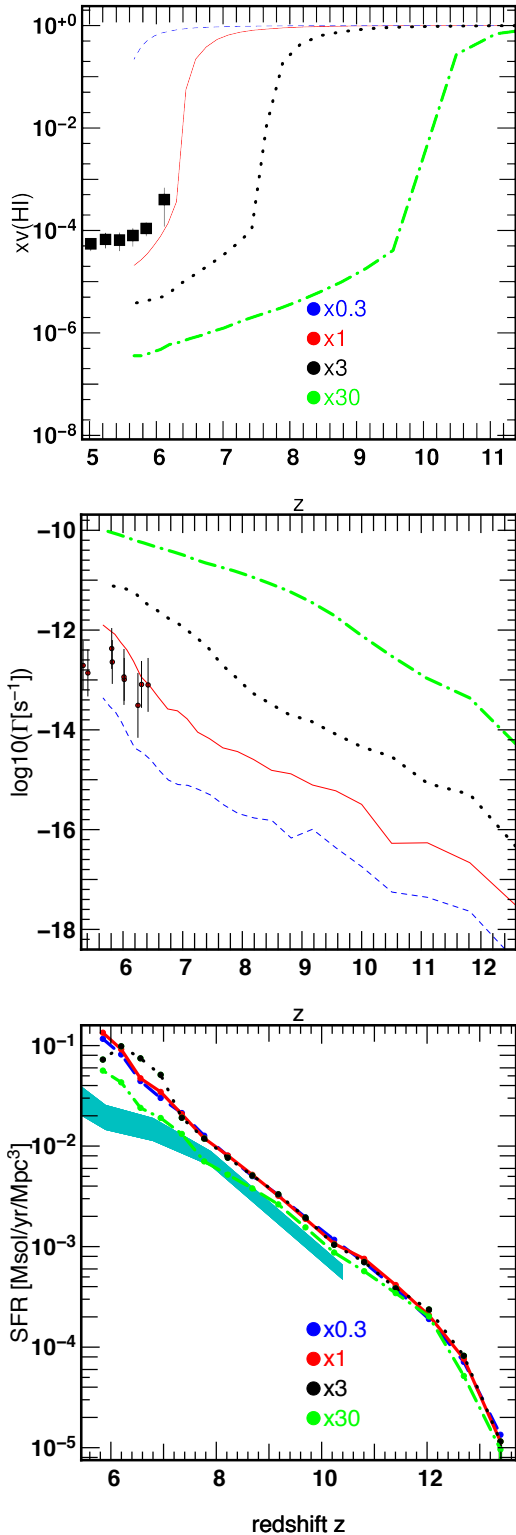


Figure 18. Global evolutions of the average volume weighted neutral fraction (top), photoionization rate (middle) and star formation history (bottom) in 4 Mpc-128³ reionization simulations. The fiducial model is shown in solid-red (labeled as having an X1 emissivity) while simulations with different emissivities are shown in dashed-blue (with an emissivity equals to 30% the fiducial one, X0.3), dotted-black (X3) and dash-dotted-green (X30). Observational constraints from Fan et al. (2006) (Top panel, squares), Calverley et al. (2011) (middle panel, points) and Bouwens et al. (2014) (bottom panel, blue shaded area) are also given.

icant scatter is obtained. If all the models are considered, a clear trend can be noted : the dimmest models (X0.3, X1) share the same global behavior (or qualitative functional form) even though the fiducial model presents baryon poorer low mass halos. Meanwhile, the brightest models can produce baryon fractions 10 times smaller than the fiducial case with a different functional relation between the baryon fraction and the halo mass. The impact of radiation on this quantity seems therefore well established in this series of models, where brighter sources have a strong impact on the gas within shallow potentials.

Interestingly, this impact does not directly translate into a modified SFR inside the halos (see Fig. 20). Again the scatter is quite important and finite mass effects can be seen in halos with small formation rates and the interpretation can therefore be difficult. Still, it appears that the 3 dimmest models (X0.3, the fiducial X1 and X3) are not significantly different and present the same mass dependence of the star formation rate within their halos. Since we found that the global baryon quantity is indeed affected, it seems to imply that the *star forming* baryons are unaffected by the source emissivity and the presence of radiation. Only the most extreme case of source emission shows a significant dip in the SFR of low mass halos : in our simplistic model of star formation, a certain level of gas depletion must be achieved to impact the production of stellar particles.

As a final note, we present in Fig. 21 the halo baryon fraction in the 4 models at the same cosmic average ionization fraction, $x = 0.75$. Of course this level of ionization is achieved at high redshift (~ 10) for the brightest model and corresponds to the last snapshot at $z = 5.5$ for the dimmest one. It can be easily seen that the baryon fraction mass distribution is essentially identical in the 4 models, taken at 4 different redshifts but at the same ionization level. It could hint that an essential ingredient of the baryon depletion is not only the source intensity but also the exposition duration to the UV background. In the previous analysis at $z = 5.5$ not only the brightest model contains the brightest sources but it also provided the longest duration over which halos are in an optically thin Universe since such models provide an early reionization. Conversely, dimmer models produce a late reionization and therefore a shorter exposition duration to the UV flux in a transparent Universe. It could impact the baryon fraction in low mass halo measured at a given time. At a given average ionization fraction, we somehow get rid of the scatter in flux exposition and look at halos from different simulations at a similar stage of their 'Universe' ionization history, with a similar structure for the UV field. And indeed, in our model, it significantly reduces the differences observed previously.

Let us recall that several important ingredients are missing in our models like the inclusion of supernovae feedback, which may enhance the SFR suppression in low-mass halos or the presence of H_2 , which fraction can greatly differ from the baryon fraction. Hence the results presented in this section indicates in a qualitative manner that **EMMA** is able to handle cosmological reionization simulations. Further investigations and implementations are necessary to quantitatively assess the subjects discussed here.

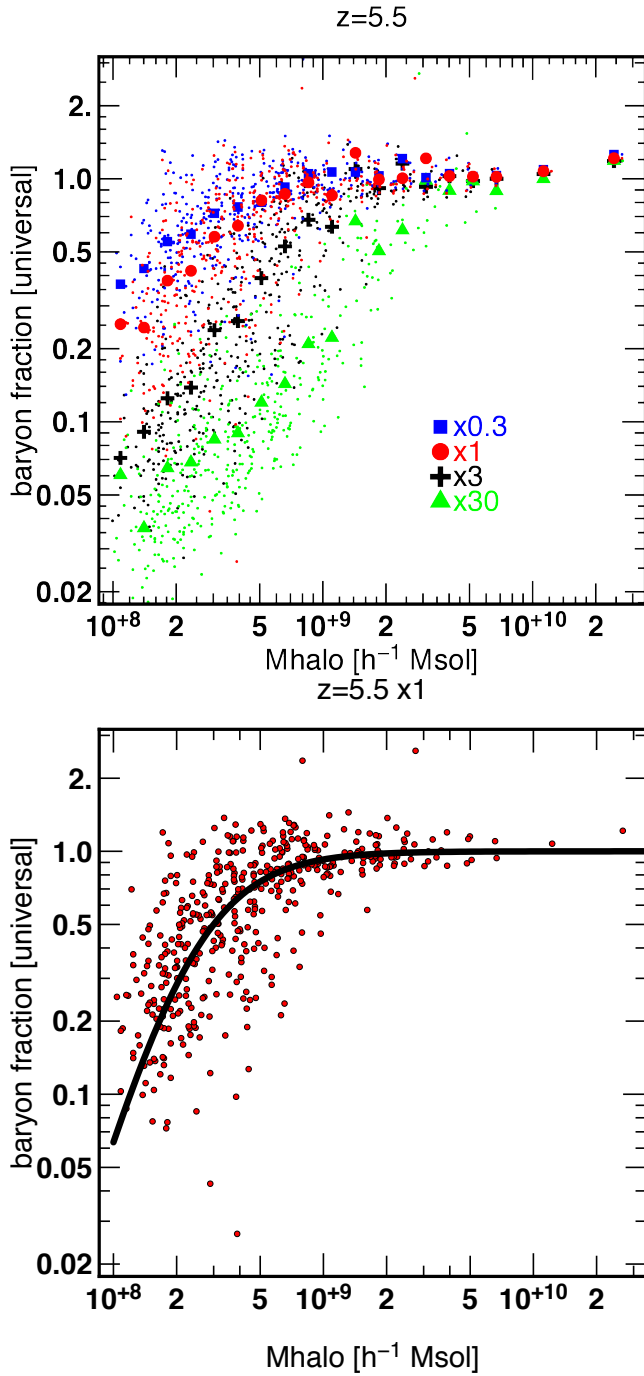


Figure 19. Top: Baryon fraction in DM halos as a function of their mass at $z=5.5$ computed in the 4 models of emissivities. Small dots stand for the values for each individual halo whereas large symbols stand for the average baryon fraction within a bin of halo mass. Bottom: the same quantity but for the fiducial model only (dots) compared to the Okamoto et al. (2008) fit.

7 PERFORMANCES

7.1 Preamble

As a closing chapter to this description of EMMA, we now discuss the performances of the code. As shown hereafter, the comparison of performances on different architectures is a complex matter as it depends on how architecture-

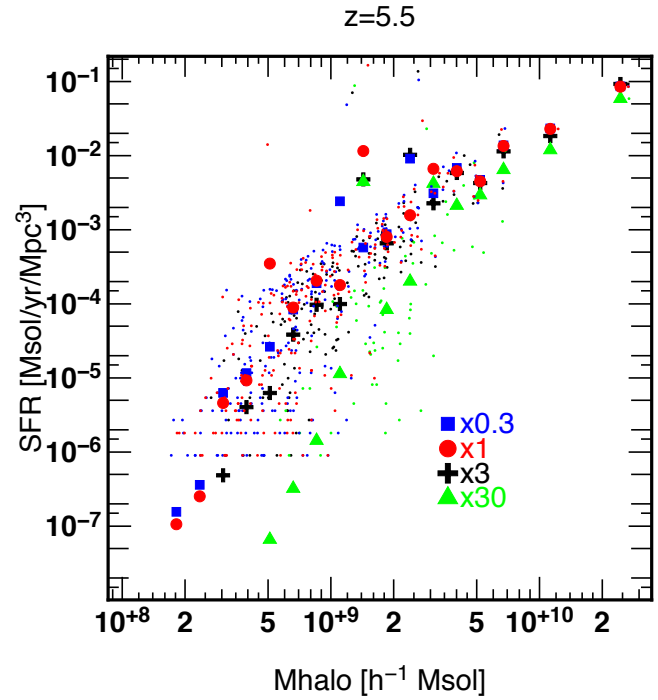


Figure 20. The instantaneous star formation rate per DM halo mass for the 4 emissivity models. Small dots stand for the values for each individual halo whereas large symbols stand for the average baryon fraction within a bin of halo mass.

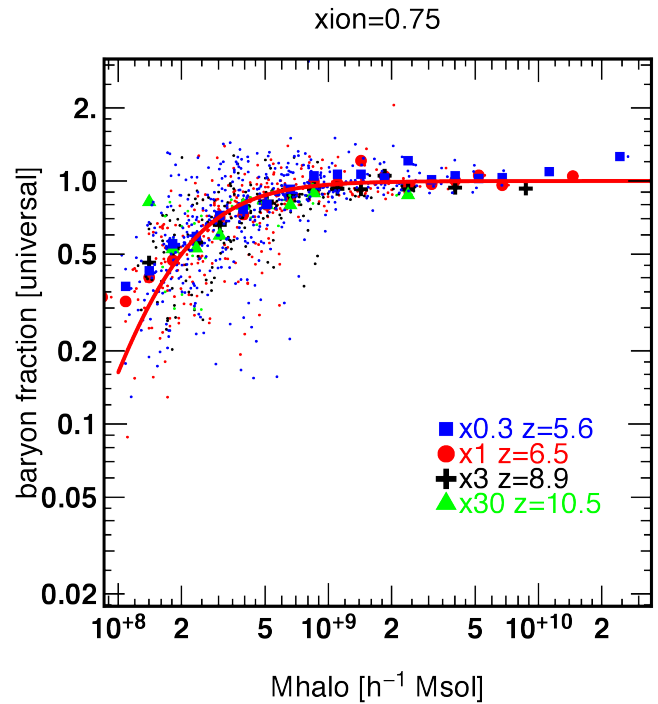


Figure 21. The baryon fraction as a function of the halo mass for the 4 different models measured at the same ionization fraction $x = 0.75$. The corresponding redshifts are given in the labels. The red line stand for the Okamoto et al. (2008) fit taken at the redshift of the fiducial model (X1).

dependent optimizations are implemented. Such a comparison also depends on the context it has been performed : as such it will evolve in time (as hardware improves for instance) or in 'space' (from one computer to another one at a given time). We nevertheless think that the following sections will shed some light on how the code behave and how these behavior can significantly vary depending on the compilers or the architecture. More generally it is also an opportunity to demonstrate that codes performances should be carefully considered, and not only for **EMMA**.

In the following sections, the calculations involved a 4 Mpc/h - 128^3 cosmological simulation with full physics and the same parameters and simple star formation recipe as the ones described in section 6.6.1. They used the CRTA approximation with a $c/10$ speed-of-light, which should not affect the discussion on raw performance and scaling issues. We compared 3 types of **EMMA** binaries on the Curie-CCRT supercomputer hybrid nodes, compiled using single precision arithmetic:

- a GPU binary produced by the NVCC compiler from the CUDA 5.5 SDK with O2 optimization level, called *gpu-O2*. This version runs the vectorized physical engines on M2090 Nvidia GPU devices while still relying on a single core to perform the other tasks, such as AMR logistics, vectorization, particles operations, etc... In the case of a multi-GPU run, each MPI process is attached to a single CPU core associated to a single distinct GPU.
- a CPU binary produced by GCC 4.4.7, called *gcc-O2* hereafter. This version fully runs on 2.7 GHz Sandybridge Westmere processors and uses the standard O2 optimization level.
- a CPU binary produced by ICC 14.0.3. with the same O2 optimization level, called *icc-O2* hereafter. Such **EMMA** binaries are usually faster than the ones provided by GCC by a factor close to 4 : this difference is essentially the result of optimizations on floating point operations that are enabled by default. Such optimizations can be disabled by setting an additional `fp-model=strict` flag and produce **EMMA** binaries with reduced performances at the level of the ones produced by gcc (not shown here).

ICC is available in most supercomputing and institutional facilities. GCC on the other hand is widely distributed and could be the only option on small configurations (e.g. on laptops, desktop machines or local shared memory calculators). Since they produce binaries with different performances, the resulting GPU acceleration will also depend on the CPU-version taken as a reference.

Comparisons of GPUs and CPUs are done by considering one graphics device against one CPU *core*. Obviously, it biases performances in favor of GPUs which are essentially parallel devices. Nevertheless, we argue that it is the simplest way to do the comparison, since a given GPU can be associated with a variety of different CPU nodes with different core numbers. However some care must be taken when considering acceleration rates. If an acceleration factor of x80 is found, it should be seen as considerable since 80 CPU cores per GPU is already a significant configuration and codes usually don't follow strong scaling laws at such levels of acceleration (i.e. an x80 acceleration cannot be obtained using 80 cores). On the other hand if an x4 acceleration factor is found, it should be considered as low

since 4 cores are easily obtained and x4 strong scaling factors can usually be achieved. In the case of Curie hybrid nodes, 1 GPU is associated with 4 cores but other configurations exist (e.g. Titan-ORNL associates 16 cores with 1 GPU).

7.2 Computing time consumption

Fig 22 presents the time spent by a calculation to reach a given expansion factor in a cosmological simulation. Whichever code version is considered, two major phases can be distinguished : for an expansion factor $a = (1+z)^{-1} < 0.065$ the code achieves a stable regime with small and regular time steps (given by the slope of the Fig. 22 curves). At this stage, no source has been created yet and no light has to be propagated : the radiative engine (which also includes thermo-chemistry modules such as cooling processes) is not limited by the CFL condition and is called once per dynamical time step. Furthermore, non linearities are small and AMR has not been deployed yet, hence the work per coarse cell is naturally close to balance. At $a \sim 0.065$ the first source appears and radiative transport must be computed while satisfying the stringent CFL condition. The number of RT calls per dynamical time step increases to typical levels of 150 calls per step. In Fig 22 the time spent increases by orders of magnitude with a much greater slope, i.e. a much greater time spent per time step. This contribution of RT to the computing time is further emphasized by the dashed blue line in Fig. 22 which stands for the RT-only contribution in the *gpu-O2* calculation (similar curves are obtained for the CPUs calculation albeit not shown here): clearly the dramatic increase in the computing time is driven by this specific module.

In the same plot, solid lines stand for the computing time required for **EMMA** to reach a given expansion factor using a M2090 Nvidia GPU Device with *gpu-O2* (black dashed line) and using a single CPU core with *gcc-O2* (black solid line) and *icc-O2* (black dotted line) binaries. Comparing these different versions of **EMMA**, it can be seen in Fig. 22 that for the *gpu-O2* version $a = 0.07$ is achieved in 50 minutes, whereas 16 hours are required for the *gcc-O2* version, providing an x16.9 acceleration factor. In the pre-source regime (for $a < 0.065$) this acceleration factor drops to x6.4: in this regime the contribution of the radiative transfer engine is much smaller and so is the level of potential acceleration. In the very first stages of the calculation this acceleration rate even drops further (to factor close to x4) as the cooling induced by dynamical effects is small and hence the need for associated calculations that could have benefited from hardware acceleration. If the GPU version is compared to the *icc-O2* run, the maximal acceleration rate of the *gpu-O2* code drops to x3.9. Clearly the removal of strict value-safe floating point operations (which allows greater optimization from the CPU compiler) results in a more competitive CPU code performance-wise. Moreover, it should be noted that the current comparisons deal with a GPU device against a single CPU core as we argued that it provides the simplest mean of comparison. However cores are usually part of multi-core nodes, connected to one or two GPU devices. Hence an acceleration rate of a few can be seen as not sufficient if it does not exceed the core per GPU ratio. For instance we show in Fig. 22 the time required for the 4 cores of hybrid Curie *node* to run the same test (symbols). As can

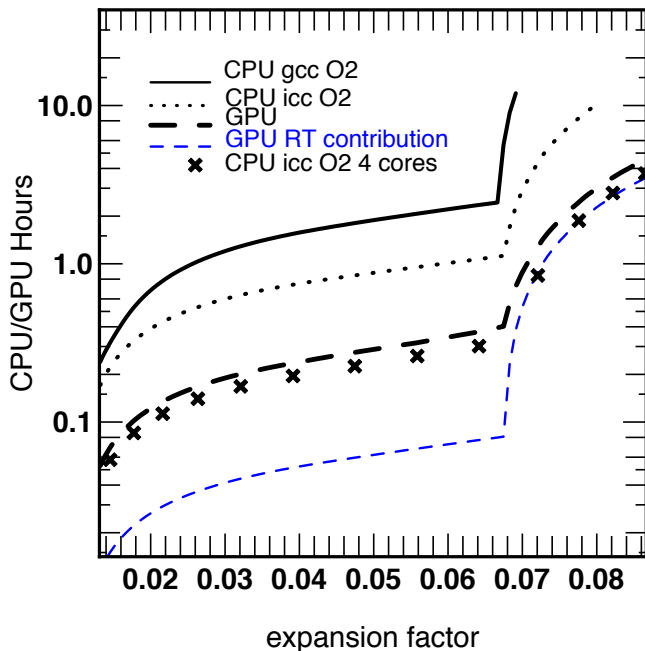


Figure 22. Comparison of the cumulative time spent to reach a given expansion factor for a 4 Mpc/h-128³ cosmological simulation of the reionization. Times are given for a single computing device (i.e. 1 GPU or 1 CPU core). The thick black dashed line stands for the GPU run performed on a M2090 Nvidia GPU whereas the thin dashed blue line stands for the contribution of radiative transfer to this cost. The black solid (resp. dotted) line stands for a single CPU core (2.7 GHz Sandybridge Westemere) using the *gcc-O2* (resp. *icc-O2*) binary. The symbols stand for a 4-core CPU calculation using *icc-O2* on a Curie node.

be seen here, the strong scaling behaviour of EMMA is sufficient to further improve the CPU-consumption by a factor of almost 4 and the parallel *icc-O2* binary slightly outperforms the *gpu-O2* performances. As described in the next subsection, these diverse performances result from the impact of CPU optimization and GPU acceleration that are not uniformly distributed among the different modules.

7.3 Detailed computing cost breakdown

For the same experiments, Fig. 23 presents the cumulative time spent in the 3 principal modules (Poisson, Hydro and Radiative solver) of EMMA as a function of the number of successive calls made to these modules. Solid lines stand for the single core CPU calculation obtained with *gcc-O2*, dotted lines stand for single core CPU calculations produced by *icc-O2* and dashed lines stand for the GPU-driven experiments.

Focusing first on the *gcc-O2* results, it appears clearly that hydro and RT calculations dominate the overall time budget of EMMA. Unsurprisingly, the Poisson solver only contributes marginally to the overall cost : first, the amount of calculation involved in this stage is small compared to the complex hydrodynamical solvers or thermo-chemistry calculations. Second, it relies on an iterative solver, where the solution does not evolve quickly from a time step to another or only in a few hyper-refined cells, ensuring a rapid convergence and hence a low computational cost. It can also

	HYD VT	HYD Cal	RT VT	RT Cal
<i>icc-O2</i>	1.2	26.4	0.9	13.1
<i>gpu-O2</i>	1.76	0.69	1.66	0.85

Table 1. Typical time spent (in seconds) in the vectorization+transfer operations (VT) and in calculations (Cal) for the hydrodynamics (HYD) and radiative transfer modules (RT). Times are given for time step #10 of the benchmark simulation described in Sec. 7, corresponding to a regime without sources and without AMR.

be noted that the hydrodynamics are the dominant stage at early times, being overtaken by RT only as thermo-chemical computations start to contribute and obviously at later time when the CRTA approximation execute ~ 150 RT calls per hydro call. This effect due to CRTA is also evident in the number of RT calls which is much greater than the identical number of hydro and Poisson calls.

Looking at the performance of the GPU -driven binaries *gpu-O2*, the time spent in the hydrodynamics and the RT is reduced to the levels of the Poisson solver : compared to *gcc-O2* the RT module is accelerated by a factor x32 and the hydrodynamics by a factor x14 . Interestingly we could not achieve any acceleration with the Poisson solver on GPU architecture. The reason is the poor computation/transfer ratio for the Poisson solver: our measurements show that gathering the data from the AMR to vector-like structure on CPU takes $\sim 75\%$ of the time required by the Poisson solver: the room for acceleration is therefore extremely small whereas in hydro and RT this gathering stage only represents $\sim 5 - 10\%$ of the computation. In general, the acceleration can be efficient on computation-dominated modules, and in our implementation the Poisson iterative solver does not belong to this family of functions and represents therefore an intrinsic limit to EMMA performances on GPUs.

Finally, dotted lines show the cumulative time per calls of a given module but using *icc-O2*. No differences can be noted for the hydro and Poisson solver compared to the timings obtained from *gcc-O2*, but the time spent into the radiative transfer module is greatly reduced. It is easily explained by the important contribution of non trivial mathematical operations in cooling rates, cross-sections, ionization rates, etc... present in the thermo-chemistry operations handled by the RT module. Such operations have a clear benefit from the optimizations made by the compiler. This is not the case for hydrodynamics : even though a MUSCL scheme involves a great number of operations, they essentially rely on simple arithmetic operations, which are less prone to optimizations. For hydrodynamics, *gpu-O2* still provides a x12 acceleration compared to *icc-O2* but RT acceleration rate drops to x5.5 : since it is the dominant process, it strongly affects the overall GPU acceleration.

The current limiting factor of GPU performance is the cost of vectorization and data transfer to and from the device. In fact, the near identical floor performance obtained by the three GPU modules is due to the irreducible cost of these operations. In Tab. 1, we list the time consumption for the vectorization-transfer stages as well as for the actual computations, measured in a typical early-stage step and for *icc-O2* and *gpu-O2* binaries. Poisson Solver results are not

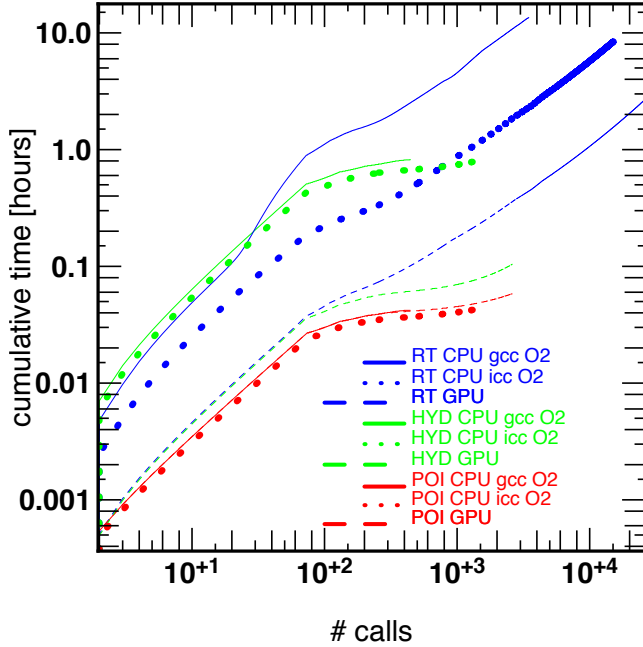


Figure 23. Cumulative time spent in the Poisson (red), hydro (green) and radiative transfer (blue) physical engines as a function of number calls. Solid lines stand for single core CPU calculations on a 4Mpc/128³ simulation using *gcc-O2*, dotted lines stand for single CPU core calculations using *icc-O2* and dashed line for calculations driven by a single M2090 GPU using *gpu-O2*.

discussed as they are already dominated by vectorization on CPU. We find that for the hydrodynamics and RT the cost of these operations is close to 70% on GPUs (see Tab. 1) whereas they contribute to less than 10% of the CPU calculation : the acceleration potential of *calculations* is thus almost fully exhausted. It can also be noted from Tab. 1 that these vectorization/transfer steps are actually more expensive on GPU, because of the additional transfer of data from the CPU host to the GPU : the cost of transfer is broadly equivalent to the vectorization. It doubles the time spent in non-calculations operations which end up dominating the cost of the hydrodynamical and radiative transfer modules.

7.4 Parallel scaling

Fig. 24 shows the scaling properties for **EMMA**, where a fix load per process is chosen and the number of process is increased, thus increasing the volume and total number of coarse cells or particles handled by the code. For the *gpu-O2* and *gcc-O2* scaling measurements, we stick to a 4Mpc - 128x128x128 coarse cells per process load, similar to the one used above. For the *icc-O2* Intel CPU scaling we varied the load per core and used non cubical sub-domains. Configurations from 1 to 256 GPUs and from 1 to 2048 cores have been used. Times were measured in the initial stages of a cosmological run, during the first 20 steps. At these early stages, non linearities are weak and AMR is not triggered: load imbalance is minimal and therefore allows a better estimation of parallelism-induced deviations. It also corresponds to the epoch where the acceleration is the weakest but it should not affect our conclusions regarding the scaling abilities of the code. Clearly the (weak) scaling trends are satisfying with

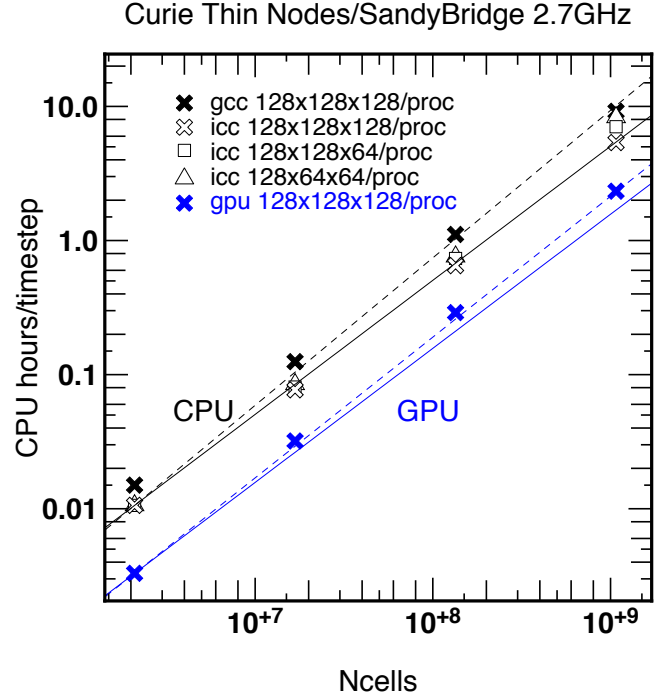


Figure 24. Empty symbols: scaling curves for different loads per process configurations using the *icc-O2* binary, given as the CPU hours per time step as a function of the total number of coarse cells. The black solid line represents the perfect CPU scaling (i.e. $t \sim N_c^\alpha$ with $\alpha = 1$) while the black dashed line stands for the worst scaling law obtained here with $\alpha = 1.1$. Filled black symbols stand for the same measurements but made with the *gcc-O2* binary and a constant load of 128³ per process. Filled blue symbols stand for the same measurements on GPUs (using *gpu-O2* binaries) and a constant load of 128³ per process. The blue solid line stands for the perfect scaling expected for GPU whereas the dashed blue one stands for the one actually measured with $\alpha = 1.05$.

a CPU/GPU computing time that scales as $t \sim N^{[1-1.1]}$ where N is the number of cells. In fact on CPU the scaling is almost perfect in configurations with a 128x128x128 load per process and drifts away for smaller problems per process : it is a standard strong scaling issue, where a smaller local load increases the weight of parallelism overheads. Furthermore, as the sub-domains become non-cubical the number of neighbors and therefore the amount of communication varies from one process to another if a Peano-Hilbert segmentation is used, leading to a small unbalance of communications between processes. Nevertheless, the scaling of **EMMA** running on multiple CPUs and GPUs seems satisfying, aside from load-balance issues that will inevitably arise later on as structures will emerge. We plan to implement balancing procedures in forthcoming developments.

7.5 Discussion

Overall, the performance achieved by GPU driven calculations is promising and acceleration rates greater up to x16.9 can be obtained, but such rates should be discussed as performance comparison is a complex matter. First the large acceleration rates are obtained in the regime where the CRTA is used and effective (i.e. after the first star has appeared):

the CRTA is somewhat designed to favor hardware accelerators as it increases the weight of pure and heavy calculation on the overall budget of **EMMA**. Furthermore it remains an approximation with a coarse description of radiative transport, hence it stands as a lower order approximation compared to the standard AMR coupling of radiation to matter. Note that even in the CRTA regime this standard coupling is naturally enforced in the pre-source stages, and we demonstrated that only moderate acceleration is achieved in this regime (x6.4). Second, we also demonstrated that properly optimized CPU binaries can be only a few times slower than GPUs. The overall code acceleration rate drops then to x3.9 even in the CRTA dominated regime, as can be seen by examining Fig. 22. Finally the performance gap can further be reduced by increasing the number of CPU cores used or by using full nodes capacities.

It should be noted that these acceleration are global ones, i.e. they rely on global timings of **EMMA** where a significant number of modules remain to be ported on GPUs. This is for instance the case for particles-related operations which are currently only handled by the CPU. Even if they are sub-dominant, optimizing these tasks or porting them for GPU architecture could provide a moderate additional acceleration.

However, the most obvious way to increase the GPU acceleration is to reduce the cost of vectorization and data transfer to the device. The data transfer bottleneck is expected to evolve naturally as new standards are being developed to increase the CPU to GPU bandwidth² or by using architectures such as AMD's Accelerated Processor Units (APUs) where CPU and GPU share the same memory, thus nullifying the cost of transfer. It should be noted that thanks to the vectorization strategy of **EMMA**, future ports on new architectures should be of limited complexity. Regarding the cost of the vectorization (driven by gather/scatter operations), let us note that this operation is currently purely sequential and dealt with by the CPU. Gather/scatter operations could in principle be parallelized to reduce its imprint on the overall costs and to lower the intrinsic floor that limits the GPU performance. Such parallelization is limited by concurrent memory accesses but current CPU architectures designed with Non Uniform Memory Access (NUMA) should in principle alleviate this issue. Another option would be to deport the vectorization on the parallel computing device (a GPU in our case) : performance gains are expected to be limited, since gather/scatter operations rely on non 'gpu-friendly' tree-walk operations, but even a weak acceleration of the vectorization process would provide a welcome boost to the calculation acceleration.

The tests presented here were made on CPU and GPU architectures of the same generation (M2090+2.7 GHz Westmere on Curie), but newer hardware is and will be available and similar tests will be necessary to reassess the results shown here. However preliminary tests on more recent devices show no significant improvement in performance (see appendix B). It does not come as a complete surprise since our GPU calculations are currently limited by vectorization+transfer and the more recent hardware does not provide significant progresses on these aspects.

On a broader perspective, the results presented here seems to make a strong case for a full usage of hybrid installations, where **EMMA** would distribute its different tasks/modules simultaneously on the different types of hardware (multicore CPU, GPU) available on a node. For instance gather/scatter operation on a CPU represents typically 5-10% of the time spent in a physics engine, hence there is room for potential acceleration on a multi-core node through OpenMP directives, probably of a factor of a few, making it competitive with current GPU accelerations. For instance, multiple tasks could be done in parallel such as e.g. thermo-chemistry on multi-core CPU and radiative transport on GPU and current performances could be increased by an overall factor of 2. Of course, these estimates need to be confirmed by experiences.

8 CONCLUSIONS

EMMA is a cosmological simulation code which handles simultaneously gravity, hydrodynamics and radiative transfer on an adaptive grid that can be refined on the fly (AMR). Written in C, this code is parallel (via the MPI protocol) and can deploy its physics modules on graphics processing units using CUDA. Designed for the study of the reionization epoch, **EMMA** is nevertheless a versatile code for structure formation.

The code passed a variety of test cases and can confidently produce accurate and relevant simulations. A first comparison of cosmological reionization simulations with different source parameters presents the expected qualitative behaviour of the physics at play.

EMMA has been tested in a wide variety of parallel configurations and it demonstrates satisfying scaling properties. It is able to use graphics processing units (GPUs) to accelerate hydrodynamics and radiative transfer calculations. Depending on the optimizations and the compilers used to generate the CPU reference, global GPU acceleration factors between x3.9 and x16.9 can be obtained. Vectorization and transfer operations currently prevent better GPU performances and we expect that future optimizations and hardware evolution will lead to greater accelerations. Overall we demonstrate that **EMMA** is able to cope efficiently with a variety of hardware.

Aside from optimization to improve the code performance and GPU-driven acceleration factors, additional features will be included into **EMMA** in a near future. Among them, star formation and supernovae feedback is a major priority as it is an essential ingredient for galaxy formation theories and models. Their implementation is on the way and will be described in a forth coming paper. Molecular chemistry is envisioned too as it is physically relevant to understand the formation of the first and smallest objects during the reionization epoch (like mini-halos with $M < 10^7 M_\odot$) and also numerically interesting as such calculations can be easily accelerated. Full documentation of the code is also on the way in order to publicly release **EMMA** on a finite, maybe short, term.

² <http://www.nvidia.com/object/nvlink.html>

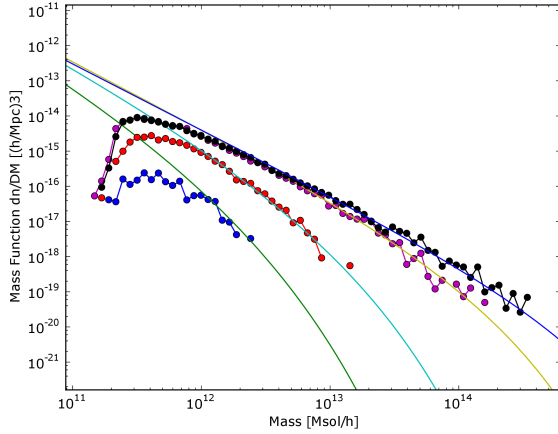


Figure A1. The halo mass function of a pure dark matter $100h^{-1}\text{Mpc}/256^3$ cosmological run. Dots stand for the mass functions measured in the simulation at $z=5.4, 3.3, 1.0$ and 0.0 (from bottom to top). Lines stand for the Sheth et al. (2001) expression for the halo mass function.

ACKNOWLEDGMENTS

We are grateful to B. Semelin, N. Gillet, J. Blaizot, J. Rosdahl, C. Pichon, R. Teyssier, T. Stranex, P. Shapiro for discussions over the years that provided the basis for the **EMMA** code. We are also grateful to the anonymous referee who helped to improve the quality of the article. This work has been supported by the ANR grants ANR-12-JS05-0001 (**EMMA**), ANR-14-CE33-0016-03 (**ORAGE**) and ANR-09-BLAN-0030 (**LIDAU**). This research used resources of the Oak Ridge Leadership Computing Facility (INCITE 2013 Award AST031, INCITE Prep project AST105), which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725, from the Meso-Centre de l'Universite de Strasbourg and from the Centre de Calcul Recherche et Technologie-CCRT (Curie-CPU & Curie-GPU, DARI Grant 2015047393). The authors thank D. Munro for freely distributing his Yorick³ programming language and its yorick-gl extension.

APPENDIX A: ADDITIONAL TESTS ON COSMOLOGICAL SIMULATIONS

A1 Mass function in a pure DM simulations

First, we try to recover the halo mass function in pure dark matter cosmological simulations. Initial conditions were produced with the MPGratic package (Prunet et al. 2008) for a $100h^{-1}$ comoving Mpc box sampled with 256^3 particles. The gravitational potential is computed on a 256^3 coarse grid ($\ell = 8$) and refinement up to $\ell = 12$ is triggered in a quasi-Lagrangian manner when a cell contains more than 8 particles. The spatial resolution is equivalent to a 4096^3 grid. Cosmological parameters were taken from Planck Collaboration et al. (2013) (setting $\Omega_m = \Omega_c$) and used as inputs to the Eisenstein & Hu (1998) transfer function. The halos

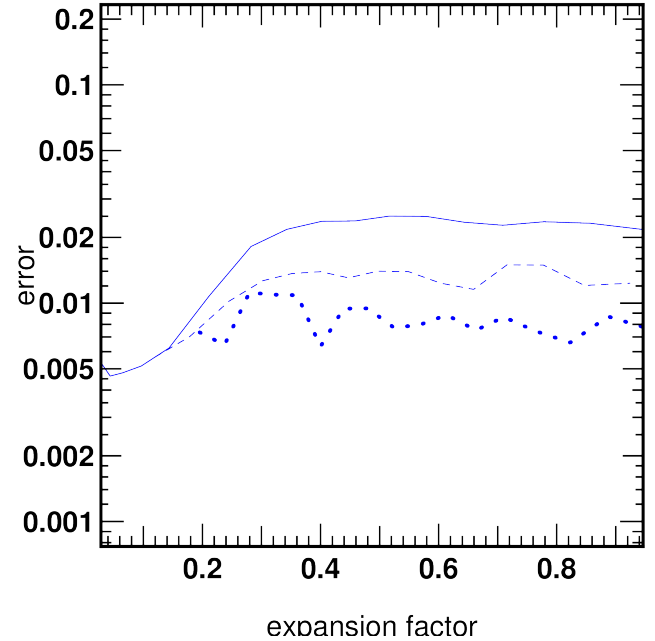


Figure A2. Error on cosmic energy variation as defined in Eq. A1 for three 100 Mpc/h adiabatic DM+gas cosmological simulations with 128^3 coarse resolutions ($\ell = 7$). From top to bottom a simulation without AMR (solid), with $\ell_{\text{max}} = 8$ (dashed) and $\ell_{\text{max}} = 10$ (dotted).

were detected using the HOP halo finder (Eisenstein & Hut 1998) and their mass function is compared to the formulation of Sheth et al. (2001). Only halos with a number of particles greater than 10 particles were kept, corresponding to a minimal mass of $5.2 \times 10^{10} M_\odot$.

Fig. A1 presents the halo mass function obtained at different redshifts, directly compared to Sheth et al. (2001). A good agreement is obtained at all redshift, with massive halos kicking in only at later time as expected. On the low-mass end of the mass function, **EMMA** is complete for halos with at least ~ 500 particles and a factor of two below full completeness for halos with ~ 100 particles. These numbers are standard for such AMR codes and overall we can conclude that **EMMA** tracks correctly the assembly history of dark matter halos.

A2 Energy 'Conservation' in DM+gas adiabatic simulations

In this section, we probe the energy conservation of an adiabatic cosmological run (DM+gas). In supercomoving variables, the cosmic energy varies between expansion factors a_1 and a_2 according to

$$(\tilde{T} + \tilde{U} + \tilde{E})|_{a_1}^{a_2} = \int_{a_1}^{a_2} \frac{\tilde{U}}{a} da \quad (\text{A1})$$

for a $\gamma = 5/3$ gas. \tilde{T} , \tilde{U} and \tilde{E} stand for the total supercomoving kinetic energy, potential energy and internal gas energy. Three $100 h^{-1}\text{Mpc}$ 128^3 ($\ell = 7$) simulations with $\ell_{\text{max}} = 7, 8, 10$ were run, using the same cosmological parameters and refinement strategy as in Sec. A1. In the same spirit as Kravtsov et al. (1997), we check Eq. A1 against the change in potential energy, i.e.:

³ <http://dhmunro.github.io/yorick-doc/>

$$\text{error} = \frac{(\tilde{T} + \tilde{U} + \tilde{E})|_{a_1}^{a_2} - \int_{a_1}^{a_2} \frac{\tilde{U}}{a} da}{\tilde{U}|_{a_1}^{a_2}} \quad (\text{A2})$$

In Fig. A2, the error is shown for 3 maximum level of refinement : $\ell_{\max} = 7$ (i.e. no refinement), $\ell_{\max} = 8$ and $\ell_{\max} = 10$. The error is found to be under control at 2%, 1.2% and 0.7% respectively. One can note how the 3 tracks diverge as refinement levels are installed (e.g. $a = 0.12$ for $\ell = 8$ and $a = 0.2$ for $\ell = 9$) providing greater resolution and smaller energy drifts. Overall, these levels of error is consistent with e.g. Kravtsov et al. (1997); Teyssier (2002).

APPENDIX B: PRELIMINARY COMPARISON OF EMMA PERFORMANCES ON DIFFERENT GPU DEVICES

We briefly describe the timings obtained by EMMA on K20c devices, more recent than the M2090 GPUs available on Curie. K20c devices differ by the number and type of cores (2496 Kepler cores versus 512 Fermi cores for the M2090), core clock (706 MHz Vs 1.3GHz for the M2090), memory frequency (2.6 GHz versus 1.9 GHz for the M2090) and bandwidth (208 Gb/s versus 177 Gb/s for the M2090). In terms of single precision floating point operations, the theoretical peak performance of K20c is a factor of 2 greater than the M2090.

We ran a 4 Mpc/h cosmological simulation on a single GPU with full physics over 100 time steps, with the same settings as the one chosen in Sec. 3 and 4. Fig. B1 compares the duration of the time steps obtained from two simulations made on these two kind of devices. In both case, the timings show the same global evolution with spikes due to outputs of data and large jumps due to AMR refinement. K20 performances are marginally better than the M2090 ones, at the 10% level, despite their greater computing power. It is expected since, EMMA calculations on GPU are already dominated on M2090 devices by gather/scatter and host to device transfer operations. Future devices with greater bandwidth could improve the situation, but in its current state EMMA does not really benefit from the greater computing power of more recent hardware.

REFERENCES

- Aubert D., Amini M., David R., 2009, *Lecture Notes in Computer Science*, 5544, 874
Aubert D., Teyssier R., 2008, *MNRAS*, 387, 295
Aubert D., Teyssier R., 2010, *ApJ*, 724, 244
Baek S., Di Matteo P., Semelin B., Combes F., Revaz Y., 2009, *A& A*, 495, 389
Baek S., Semelin B., Di Matteo P., Revaz Y., Combes F., 2010, *A& A*, 523, A4
Barkana R., Loeb A., 2001, *Phys. Rep.*, 349, 125
Bertschinger E., 1985, *ApJS*, 58, 39
Bouwens R. J., Illingworth G. D., Oesch P. A., Trenti M., Labbe' I., Bradley L., Carollo M., van Dokkum P. G., Gonzalez V., Holwerda B., Franx M., Spitler L., Smit R., Magee D., 2014, *ArXiv e-prints*
Calverley A. P., Becker G. D., Haehnelt M. G., Bolton J. S., 2011, *MNRAS*, 412, 2543

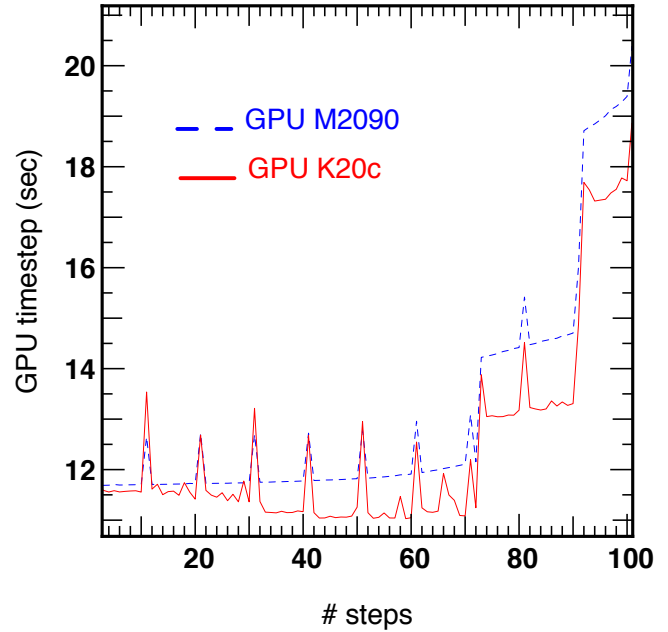


Figure B1. Comparison of the time step duration on two different kind of GPU devices. Measures were taken on the first 100 steps of a 4 Mpc/h cosmological simulation with full physics with parameters similar to the runs described in Sec. 3. The blue dashed line stands for M2090 timings, the red solid line stand for timings on the more recent K20. The spikes seen in both curve are due to input/outputs operations.

- Chardin J., Aubert D., Ocvirk P., 2012, *A& A*, 548, A9
Ciardi B., Stoehr F., White S. D. M., 2003, *MNRAS*, 343, 1101
Dubinski J., 1996, *New Astronomy*, 1, 133
Dubois Y., Teyssier R., 2008, *A& A*, 477, 79
Eisenstein D. J., Hu W., 1998, *ApJ*, 496, 605
Eisenstein D. J., Hut P., 1998, *ApJ*, 498, 137
Fan X., Strauss M. A., Becker R. H., White R. L., Gunn J. E., Knapp G. R., Richards G. T., Schneider D. P., Brinkmann J., Fukugita M., 2006, *AJ*, 132, 117
Finlator K., Davé R., Özel F., 2011, *ApJ*, 743, 169
Gnedin N. Y., 2000, *ApJ*, 542, 535
Gnedin N. Y., 2014, *ApJ*, 793, 29
Gnedin N. Y., Abel T., 2001, *New Astronomy*, 6, 437
González M., Audit E., Huynh P., 2007, *A& A*, 464, 429
Hockney R. W., Eastwood J. W., 1981, *Computer Simulation Using Particles*
Hoeft M., Yepes G., Gottlöber S., Springel V., 2006, *MNRAS*, 371, 401
Hui L., Gnedin N. Y., 1997, *MNRAS*, 292, 27
Iliev I. T., Ciardi B., Alvarez M. A., Maselli A., Ferrara A., Gnedin N. Y., Mellema G., Nakamoto T., Norman M. L., Razoumov A. O., Rijkhorst E.-J., Ritzerveld J., Shapiro P. R., Susa H., Umemura M., Whalen D. J., 2006, *MNRAS*, 371, 1057
Iliev I. T., Whalen D., Mellema G., Ahn K., Baek S., Gnedin N. Y., Kravtsov A. V., Norman M., Raicevic M., Reynolds D. R., Sato D., Shapiro P. R., Semelin B., Smidt J., Susa H., Theuns T., Umemura M., 2009, *MNRAS*, 400, 1283
Katz N., Weinberg D. H., Hernquist L., 1996, *ApJS*, 105, 19

- Kay S. T., Pearce F. R., Frenk C. S., Jenkins A., 2002, MNRAS, 330, 113
- Khokhlov A., 1998, Journal of Computational Physics, 143, 519
- Kravtsov A. V., Klypin A. A., Khokhlov A. M., 1997, ApJS, 111, 73
- Levermore C. D., 1984, JQSRT, 31, 149
- Martel H., Shapiro P. R., 1998, MNRAS, 297, 467
- McQuinn M., Lidz A., Zahn O., Dutta S., Hernquist L., Zaldarriaga M., 2007, MNRAS, 377, 1043
- Mellema G., Iliev I. T., Pen U.-L., Shapiro P. R., 2006, MNRAS, 372, 679
- Norman M. L., Reynolds D. R., So G. C., Harkness R. P., Wise J. H., 2015, ApJS, 216, 16
- Ocvirk P., Aubert D., Chardin J., Knebe A., Libeskind N., Gottlöber S., Yepes G., Hoffman Y., 2013, ApJ, 777, 51
- Ocvirk P., Gillet N., Aubert D., Knebe A., Libeskind N., Chardin J., Gottlöber S., Yepes G., Hoffman Y., 2014, ApJ, 794, 20
- Okamoto T., Gao L., Theuns T., 2008, MNRAS, 390, 920
- Paardekooper J.-P., Khochfar S., Dalla Vecchia C., 2013, MNRAS, 429, L94
- Pawlik A. H., Schaye J., Dalla Vecchia C., 2015, ArXiv e-prints
- Planck Collaboration Ade P. A. R., Aghanim N., Armitage-Caplan C., Arnaud M., Ashdown M., Atrio-Barandela F., Aumont J., Baccigalupi C., Banday A. J., et al. 2013, ArXiv e-prints
- Pontzen A., Governato F., 2012, MNRAS, 421, 3464
- Pritchard J. R., Loeb A., 2012, Reports on Progress in Physics, 75, 086901
- Prunet S., Pichon C., Aubert D., Pogosyan D., Teyssier R., Gottloeber S., 2008, ApJS, 178, 179
- Rasera Y., Teyssier R., 2006, A& A, 445, 1
- Rosdahl J., Blaizot J., Aubert D., Stranex T., Teyssier R., 2013, MNRAS, 436, 2188
- Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
- Teyssier R., 2002, A& A, 385, 337
- Theuns T., Leonard A., Efsthathiou G., Pearce F. R., Thomas P. A., 1998, MNRAS, 301, 478
- Toro E. F., 1997, Riemann solvers and numerical methods for fluid dynamics : a practical introduction. Springer, Berlin, New York
- Toro E. F., Spruce M., Speares W., 1994, Shock Waves, 4, 25
- Trac H., Cen R., 2007, ApJ, 671, 1
- Trac H. Y., Gnedin N. Y., 2011, Advanced Science Letters, 4, 228
- Warren M. S., Salmon J. K., , 1993, A Parallel Hashed Oct-Tree N-Body Algorithm
- Wise J. H., Abel T., Turk M. J., Norman M. L., Smith B. D., 2012, MNRAS, 427, 311
- Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2007, ApJ, 654, 12
- Zawada K., Semelin B., Vonlanthen P., Baek S., Revaz Y., 2014, MNRAS, 439, 1615
- Zel'dovich Y. B., 1970, A& A, 5, 84